

AD-A185 459

DEVELOPMENT OF STATISTICAL METHODS USING PREDICTIVE
INFERENCE AND ENTROPY(U) SCIENTIFIC SYSTEMS INC
CAMBRIDGE MA W E LARIMMRE MAR 86 SSI-1112

1/1

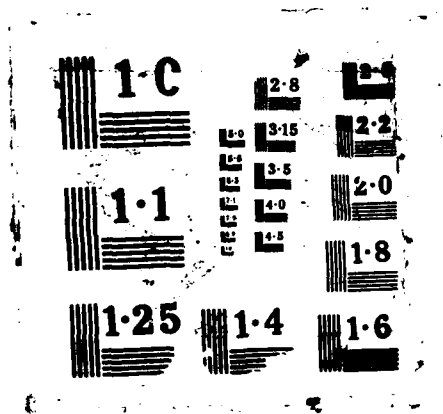
UNCLASSIFIED

AFOSR-TR-87-1336 F49620-85-C-0086

F/G 12/3

NL

END
11/2/
13/10



AD-A185 459

DTIC FILE COPY

(2)

UNCLASSIFIED

SECURITY CLASSIFICATION C

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED		1b. RESTRICTIVE MARKINGS	
2a. SECURITY CLASSIFICATION UNCLASSIFIED		3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.	
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE		4. PERFORMING ORGANIZATION REPORT NUMBER(S) CRD	
5. MONITORING ORGANIZATION REPORT NUMBER(S) AFOSR-TR- 87-1336		6a. NAME OF PERFORMING ORGANIZATION Scientific Sys. Inc.	
6b. OFFICE SYMBOL (If applicable)		7a. NAME OF MONITORING ORGANIZATION Air Force Office of Scientific Research	
6c. ADDRESS (City, State and ZIP Code) One Alewife Place 54 Cambridge Park Drive Cambridge, Mass. 02140		7b. ADDRESS (City, State and ZIP Code) Directorate of Mathematical & Information Sciences, Bolling AFB DC 20332-6448	
8a. NAME OF FUNDING/SPONSORING ORGANIZATION AFOSR		8b. OFFICE SYMBOL (If applicable) NM	
8c. ADDRESS (City, State and ZIP Code) Bldg 410 Bolling AFB DC 20332-6448		9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER F49620-85-C-0086	
10. SOURCE OF FUNDING NOS.		11. TITLE (Include Security Classification) Development of Statistical Methods using Predictive Inference and Entropy	
PROGRAM ELEMENT NO. 61102F		PROJECT NO. 2304	
TASK NO. A1		WORK UNIT NO.	
12. PERSONAL AUTHOR(S) Wallace E. Larimore		13a. TYPE OF REPORT Final	
13b. TIME COVERED FROM _____ TO _____		14. DATE OF REPORT (Yr., Mo., Day) March 1986	
15. PAGE COUNT		16. SUPPLEMENTARY NOTATION	
17. COSATI CODES		18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)	
FIELD		GROUP	
SUB. GR.			
19. ABSTRACT (Continue on reverse if necessary and identify by block number)			
In this Phase I study funded under the Small Business Innovation Research (SBIR) program, statistical methods are developed using the predictive inference and entropy approach. Previous recent research has derived entropy as the natural measure of model approximation error from the fundamental statistical principles of sufficiency and repeated sampling. In this study, the areas of nonnested multiple comparison, multivariable time series analysis, adaptive time series analysis of changing processes, and optimal small sample inference are investigated. Constrained maximum likelihood methods are developed for general nonnested multiple comparison. For the asymptotic optimality of these methods, a condition on the Fisher info. and Hessian matrices must be satisfied. Applying these results to multivariate time series analysis, lower bounds are derived for the achievable accuracy of the estimated transfer function and spectral matrices. Markov and canonical variate analysis (CVA) provide a means of numerically and statistically stable model fitting of multivariable time series, and these methods provide a basis for modeling and fitting time varying models of		(cont)	
20. DISTRIBUTION/AVAILABILITY OF ABSTRACT UNCLASSIFIED/UNLIMITED <input checked="" type="checkbox"/> SAME AS RPT. <input type="checkbox"/> DTIC USERS <input type="checkbox"/>		21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED	
22a. NAME OF RESPONSIBLE INDIVIDUAL Maj. Woodruff		22b. TELEPHONE NUMBER (Include Area Code) (202) 767-5027	
22c. OFFICE SYMBOL NM			

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

changing processes. Methods are derived for the optimal selection of data length for fitting slowly changing processes as well as for optimal selection of the data interval for detection of abrupt changes. Optimal small sample methods for multivariate analysis are studied, and entropy methods are shown to provide significant improvements in very small samples. Recommendations for Phase II research and development focus on the adaptive and nonadaptive time series analysis procedures developed in this study.

UNCLASSIFIED

AFOSR-TR- 87 - 1336

SSI-1112

Final Technical Report

**DEVELOPMENT OF STATISTICAL
METHODS USING PREDICTIVE
INFERENCE AND ENTROPY**

March 1986

by

Wallace E. Larimore

Prepared for:

**AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
Bolling Air Force Base
Washington D.C., 20332-6448**

Under Contract No. F49620-85-C-0086

**SCIENTIFIC SYSTEMS, INC.
One Alewife Place
54 CambridgePark Drive
Cambridge, Massachusetts 02140**



Accession For	
NTIS	CRA&I <input checked="" type="checkbox"/>
DTIC	TAB <input type="checkbox"/>
Unannounced <input type="checkbox"/>	
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or Special
A-1	

87 9 24 182

ABSTRACT

In this Phase I study funded under the Small Business Innovation Research (SBIR) program, statistical methods are developed using the predictive inference and entropy approach. Previous recent research has derived entropy as the natural measure of model approximation error from the fundamental statistical principles of sufficiency and repeated sampling. In this study, the areas of nonnested multiple comparison, multivariable time series analysis, adaptive time series analysis of changing processes, and optimal small sample inference are investigated. Constrained maximum likelihood methods are developed for general nonnested multiple comparison. For the asymptotic optimality of these methods, a condition on the Fisher information and Hessian matrices must be satisfied. Applying these results to multivariate time series analysis, lower bounds are derived for the achievable accuracy of the estimated transfer function and spectral matrices. Markov and canonical variate analysis (CVA) provide a means of numerically and statistically stable model fitting of multivariable time series, and these methods provide a basis for modeling and fitting time varying models of changing processes. Methods are derived for the optimal selection of data length for fitting slowly changing processes as well as for optimal selection of the data interval for detection of abrupt changes. Optimal small sample methods for multivariate analysis are studied, and entropy methods are shown to provide significant improvements in very small samples. Recommendations for Phase II research and development focus on the adaptive and nonadaptive time series analysis procedures developed in this study.

TABLE OF CONTENTS

SECTION	PAGE
1. INTRODUCTION AND OVERVIEW	1-1
2. APPROACH USING PREDICTIVE INFERENCE AND ENTROPY	2-1
2.1 Derivation of the Entropy Measure of Approximation	2-1
2.2 Use of Entropy in Statistical Inference	2-4
3. CONSTRAINED NONNESTED MULTIPLE COMPARISON OF MODELS	3-1
3.1 Constrained Maximum Likelihood Estimation	3-1
3.2 Unbiased Estimation of Entropy	3-3
3.3 Nested Tests	3-4
4. TIME SERIES ANALYSIS USING ENTROPY METHODS	4-1
4.1 Achievable Spectral Accuracy	4-1
4.2 Markovian Models of Time Series	4-5
4.3 Canonical Variate Analysis of Time Series	4-7
5. ADAPTIVE TIME SERIES ANALYSIS	5-1
5.1 Models for Changing Processes	5-2
5.2 Adaptation to Slow Variations	5-4
5.3 Adaptation to Abrupt Changes	5-7
6. SMALL SAMPLE MULTIVARIATE ANALYSIS	6-1
6.1 Bayesian Predictive Inference	6-1
6.2 Best Invariant Predictive Densities	6-3
6.3 Comparison of Entropy for Multivariate Normal	6-3
REFERENCES	R-1
APPENDIX A: Constrained Nonnested Multiple Comparison Using Predictive Inference and Entropy	A-1
APPENDIX B: Achievable Accuracy in Parametric Estimation of Multivariate Spectra	B-1

LIST OF FIGURES AND TABLES

	FIGURES	PAGE
Figure 1.	Spectral Accuracy of Different Model Fitting Procedures (lower curves) in Approximating the True Spectral Density (upper curve).	4-5
Figure 2.	Instantaneous Frequency of Signal (solid) and Interference (dashed).	5-5

	TABLES	PAGE
Table 1.	Dependence of Significance Level α on the Number n of Additional Parameters.	3-5
Table 2.	Value of Per Sample AIC for Subintervals of the Sample, for the Average of 128 Points, and for All Data.	5-6
Table 3.	Instantaneous Frequency Estimates.	5-7
Table 4.	Value of Per Sample AIC for Subintervals of the Sample with No Abrupt Change.	5-10
Table 5.	Value of Per Sample AIC for Subintervals of the Sample with an Abrupt Change in Dynamics at Sample 325.	5-10
Table 6.	Value of Per Sample AIC for Subintervals of the Sample with an Abrupt Change in Excitation Noise Variance at Sample 325.	5-11
Table 7.	Value of Per Sample AIC for Subintervals of the Sample with an Abrupt Change in State at Sample 325.	5-11
Table 8.	Expected Negative Entropy (and the Geometric Mean of the Likelihood Odds Relative to the Best Invariant).	6-4

1. INTRODUCTION AND OVERVIEW

In this study, statistical methods are developed using predictive inference and entropy. This approach to statistical inference allows the treatment of several difficult statistical problems that are not easily dealt with using traditional statistical methods. The particular statistical problems addressed are

- statistical model building involving the determination of parametric model structure and order in the general case of multiple nonnested alternatives,
- time series modeling and forecasting involving the determination of parametric model structure and order,
- adaptive time series analysis involving optimal methods for tracking slow changes as well as for detecting abrupt changes or failures,
- small sample inference for multivariate distributions of the exponential family.

A number of issues in these topics are resolved naturally in the predictive inference and entropy setting. This report provides an overview of the progress of the Phase I research with detailed technical papers included in the Appendices.

The recent interest in predictive distributions has come from several directions. Modern developments apparently begin with Jeffreys (1961, p.143) using a Bayesian approach as has much of the work following (Aitchison and Dunsmore, 1975, preface and p.39). The frequentist viewpoint taken in this proposal has been stimulation by small sample problems (Murray, 1977, 1979), model order and structure determination problems involving parametric models (Akaike, 1973, 1974), and nonnested multiple comparison problems (Larimore, 1977a, 1977b). Classical methods are conceptually ill-suited or perform poorly in practice on such problems.

In the first Chapter, the approach using predictive inference and entropy is described. The basis of this approach is the derivation of entropy from the fundamental statistical principles of sufficiency and repeated sampling in the context of the predictive inference setup as first presented in Larimore (1983a). This provides a sound theoretical foundation that was previously lacking for the use of entropy as the natural measure of the error in approximating a true future density by an estimated predictive density based upon a present sample. The generality of this entropy measure allows comparison of statistical inference methods and the derivation of more optimal inference procedures including

- general inference methods such as parametric or nonparametric methods
- exact evaluation of small sample procedures
- determination of model order or structure including the case of non-nested multiple comparison
- time series analysis including definition of optimal tracking of time varying processes and optimal detection of abrupt changes.

The generality of the predictive inference and entropy approach provides a basis for the generalization of the present statistical and predictive inference methods to more general statistical problems.

In Chapter 2, the multiple comparison of nonnested constrained models is developed. Previous developments in multiple comparison have considered largely the nested case and assume that the true model is contained in one of the models. In the present study, the case of constrained maximum likelihood estimation is considered where the true model may not be contained in any of the hypothesized models. The entropy measure provides a measure that allows the multiple comparison problem to be viewed as a model approximation problem. In this more general context the AIC procedure and generalizations of it are found to give asymptotically optimal predictive inference procedures as measured by entropy. In the nested case, these procedures reduce to the generalized likelihood ratio (GLR) test where the probability of rejection is a function of the number of additional parameters in the alternative model not contained in the null hypothesis.

Time series analysis for stationary processes are considered in Chapter 4. The entropy measure provides a direct interpretation of the achievable accuracy in estimation of the power spectrum of a process. The entropy is expressed as a squared relative error in estimating the spectrum. A generalization of this to multiple time series relates to principle components of the process cross spectral matrix. A lower bound is determined such that the expected integral of the squared relative error in spectral estimation is bounded by the number of estimated parameters divided by twice the sample size. An example of spectral estimation of an ARMA(4,3) process using spectral smoothing, Autoregressive modeling, and ARMA modeling shows the relative error in these estimation methods as dependent on the number of estimated parameters.

The topics of Markovian time series or state space models provides an approach to time series analysis that is readily computable and is easily extended to the case of changing processes. The general state space model form is developed for Markov processes. The canonical variate analysis (CVA) method gives a direct and numerically stable computational method for determining state space models from observational data. The basic computational method is the generalized singular value decomposition (SVD). This method allows for the direct determination of the

optimal model state order without the computationally intensive fitting of such models for the evaluation of model fit. Once the model state order is determined, the state space model coefficients are simply computed by regression. This method generalizes easily to changing processes.

Adaptive time series methods are developed in Chapter 5. Primarily two types of changes are considered, slowly varying changes and abrupt changes such as faults. Time varying Markov processes are developed for such changing processes. Such processes provide the hypothesized models for developing optimal tracking and detection of abrupt changes. An AIC based procedure is derived for the near optimal selection of the data length to use in model fitting. An example is given of estimating the spectrum of a time varying processes that gives results near the best previous solutions that are much more specialized. For abrupt change detection, a generalization of the AIC procedure is required since the comparison of models fitted on different data intervals is required which is not considered in the AIC formulation. Application of these methods to simulated data of abrupt changes in an ARMA(4,3) processes including jumps in the state, changes in the dynamics, and change in the variance of the excitation noise processes, demonstrates that the procedure is sensitive to the detection of these very different types of abrupt changes.

Small sample multivariate inference procedures are described in Chapter 6. Since the entropy measure gives an exact rational measure of the relative error of statistical inference procedures in small samples, it provides the bases for evaluation and development of small sample inference methods. The historical approach to predictive inference involves the derivation of a Bayesian predictive density. Although the method is Bayesian, in certain instances, the resultant predictive density has certain invariance properties which lead to an optimal predictive density in terms of the entropy measure. Another approach involves the direct solution for the optimal invariant predictive density minimizing the entropy measure. This optimal invariant procedure leads to the same predictive density as the Bayesian predictive density using a noninformative prior. These methods are compared for the multivariate normal distribution with the estimative and best normal estimation procedures.

2. APPROACH USING PREDICTIVE INFERENCE AND ENTROPY

The concepts of prediction and inference based on a set of data are very old and underlie much of the scientific method. While the scientific method has been much discussed in philosophical and qualitative terms, there has been very little in the literature from a basic statistical viewpoint. The most extensive literature appears to be that associated with predictive densities or predictive distributions (see Aitchison and Dunsmore, 1975). That approach is to a large degree Bayesian, although more recent treatments have developed a purely frequency sampling interpretation in connection with use of entropy or Kullback information. The weak point in the frequency approach was the seemingly arbitrary use of the entropy measure of model approximation error. More recently, however, the result of Larimore (1983a) has established the fundamental nature of the entropy measure based upon the statistical principles of sufficiency and repeated sampling. The entropy measure has in addition a very natural interpretation as the log relative odds in comparing two predictive densities in predicting the future sample. This gives a central role to the entropy measure. The use of the entropy measure for decisions on model order and structure was pioneered by Akaike (1973), and has been applied to many diverse statistical problems particularly in time series analysis. The justification given to the entropy measure in the Akaike approach, however, has been largely heuristic. Because of the importance of the justification of the entropy measure, the derivation and important concepts are outlined below in Section 2.1. The use of entropy in comparing model structure selection procedures and for exact small sample inference is discussed in Section 2.2. This approach to developing statistical procedures using predictive inference and entropy is then applied to the various topics in the following chapters of the report.

2.1 Derivation of the Entropy Measure of Approximation

Predictive inference involves an experimental situation with two trials, an informative trial with observations x and a predictive trial with observations y . The joint distribution of the two trials is permitted any statistical dependence and is described by a joint probability density $p(x, y)$. The objective is to choose a predictive distribution or density which, for each possible observed x , is a probability density for the future outcome y . More precisely, consider a family $p(y|x, \alpha)$ of predictive densities where the index α specifies a particular predictive distribution. For a particular choice of α , $p(y|x, \alpha)$ can be viewed as a conditional probability density of y given x for predicting the distribution of the future observation y given an observed value of x . The predictive inference problem involves selection of a criterion of fit for appraising the goodness of

approximation to the true conditional probability density $p(y|x)$ by the various predictive densities $p(y|x, \alpha)$ specified by different α . The choice of such a criterion of fit is the primary topic of this section. Negative entropy is derived as the natural measure of model approximation error for any predictive distribution.

Consider a family $C = \{p_\alpha(y|x), \alpha \in A\}$ of predictive densities for approximating the true density $p_*(y|x)$ of the predictive experiment y given the informative experiment x , where x and y are vectors of dimension K and L respectively with true joint density $p_*(x, y)$. For the predictive inference problem, a relative measure of goodness of approximation of $p_*(y|x)$ by the various $p_\alpha(y|x)$ is desired. To this end, a repeated sampling experiment is considered in which joint random samples (x_i, y_i) for $i=1, \dots, N$, are drawn repeatedly from a population with density $p_*(x, y)$. The probability density of the joint predictive experiments $Y = (y_1, \dots, y_N)$ predicted by the α -th model using $X = (x_1, \dots, x_N)$ is

$$p_\alpha(Y|X) = \prod_{i=1}^N p_\alpha(y_i|x_i) \quad (2-1)$$

The probability density for Y can be considered as indexed by the pair (α, X) . Statistical inference is considered about the true density $p_*(Y|X)$ of Y from among the family of probability densities $F = \{p_\alpha(Y|X), \alpha \in A\}$ for a fixed X .

To consider the essential statistical information about the future sample Y given by the predictive densities $p_\alpha(Y|X)$, the sufficiency of the likelihood function (Zacks, 1971, p. 61) is used. From this principle, any inferences about the family F drawn on the basis of the sample $(Y|X)$ follow from the observed values of the likelihood function $p_\alpha(Y|X)$ for $\alpha \in A$. The set of likelihood ratios formed from pairs of these likelihoods is also a sufficient statistic (Cox and Hinkley, 1974, p. 20-1, see also p. 37-9 for a discussion of likelihood and sufficiency principles).

For inference about the densities p_1 and p_2 , all of the information is contained in the likelihood ratio

$$\Lambda_N = \prod_{i=1}^N \frac{p_1(y_i|x_i)}{p_2(y_i|x_i)} \quad (2-2)$$

which has the intuitive interpretation of the relative odds of observing the data Y of the repeated predictive trials from each of the distributions p_1 and p_2 given fixed informative data X . The behavior of Λ_N as the number of repetitions becomes large is most easily seen by expressing it as

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log \Lambda_N(p_1, p_2) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \log \frac{p_1(y_i|x_i)}{p_2(y_i|x_i)}$$

$$\begin{aligned}
&= \iint p_*(y|x) \log \frac{p_1(y|x)}{p_2(y|x)} dx dy \\
&= \iint p_*(y|x) \log \frac{p_1(y|x)}{p_2(y|x)} dy p_*(x) dx
\end{aligned} \tag{2-3}$$

For a large number of repetitions, the odds will overwhelmingly favor p_1 or p_2 if the limit is respectively strictly positive or strictly negative. The preference for one distribution over the other as expressed by the likelihood ratio tends to grow exponentially with the number N of repeated trials. If (2-3) is zero, then there will be no consistent preference with large numbers of trials. Although for a finite number N of repetitions the likelihood ratio Λ_N depends upon the particular samples (X, Y) , asymptotically for large numbers of repetitions this dependence disappears.

The direct pairwise comparison of predictive densities is not necessary if the Kullback-Leibler conditional discrimination information (Kullback and Leibler, 1951; Kullback, 1959, p. 13)

$$I_{y|x}(p_*, p_a) = \int p_*(y|x) \log \frac{p_*(y|x)}{p_a(y|x)} dy \tag{2-4}$$

of p_a relative to p_* is used which is a function of x . Note that the order of p_a and p_* are not interchangeable with the latter playing the role of the truth.

The likelihood ratio (2-3) is expressed in terms of the Kullback information as

$$\lim_{N \rightarrow \infty} \frac{1}{N} \log \Lambda_N = E_x \{I_{y|x}(p_*, p_2) - I_{y|x}(p_*, p_1)\} \tag{2-5}$$

where E_x denotes expectation with respect to the true density $p_*(x)$. The criterion is thus determined as the *negative entropy*, or *negentropy* for brevity, defined as

$$R(p_*, p_a) = E_x I_{y|x}(p_*, p_a) = \int p_*(x) dx \int p_*(y|x) \log \frac{p_*(y|x)}{p_a(y|x)} dy \tag{2-6}$$

the expected Kullback conditional information of the predictive density relative to the true conditional density $p_*(y|x)$. In the repeated sampling experiment, the predictive density with the smaller negentropy relative to the true is ultimately preferred. The negentropy (2-6) thus orders the goodness of a set of predictive densities in approximating the true density. Also in comparing any two predictive densities p_1 and p_2 , the respective difference has the intuitive interpretation as the exponential rate at which the likelihood ratio diverges.

The above derivation of the entropy measure of approximation of a predictive density uses only the predictive inference setup in the repeated sampling context along with the principle of

sufficiency. The sufficiency principle is one of the few generally accepted principles in statistical inference. Various repeated sampling principles have been formulated, however the difficulty has been the choice of an evaluation criterion for comparing various sampling distributions. The entropy measure gives a criterion that is based upon basic statistical principles of inference.

2.2 The Use of Entropy in Statistical Inference

The entropy measure of error in approximating a predictive density is very general and can be applied in diverse modeling problems. In this section, some of the general model selection problems are described which include the nonnested multiple comparison problem, adaptive time series analysis of changing processes, and optimal small multivariate methods.

From the derivation of the entropy measure, it can be seen that the entropy measure has a number of very attractive features:

- It applies to completely general modeling problems including non-parametric methods.
- It applies exactly to small samples.
- Only the fundamental statistical principles of sufficiency and repeated sampling are used.
- It applies to time correlated problems such as time series model identification and tracking.
- Statistical inference can be fundamentally viewed as model approximation.

Note also that the predictive distribution can include an entire model structure-determination/parameter-estimation scheme by setting

$$p_k(y|x) = p(y, \hat{\theta}_{\hat{k}(x)}(x)) \quad (2-7)$$

where for every x , $\hat{k}(x)$ is the k minimizing a model structure determination criterion. Thus for each α , p_α can be regarded as a model fitting procedure including the choice $\hat{k}(x)$ of model structure.

The negentropy measure is entirely applicable to exact small sample inference, system identification, and detection of abrupt changes that include decisions among a multitude of parametric model structures which may be nonnested. Several predictive inference problems have been considered in the literature. Previous work has used the negative entropy measure in much more restricted formulations where

- the informative and predictive samples were assumed to be independent (Aitchison and Dunsmore (1975), Akaike (1973)) which does not include the time series forecasting problem
- the use of the negative entropy measure was considered as arbitrary (Aitchison (1975), Murray 1979)) or justified only asymptotically for large samples by heuristic arguments (Akaike (1973))
- the negative entropy measure was justified as a basis for comparing distinct parametric model structures (Akaike (1973)), but not for comparing model selection procedures which include choice of the model structure (Larimore, 1983a)
- only the case of nested structures such as autoregressive models were justified (Akaike (1973)) although there has been wide spread application of it to the general nonnested case such as ARMA models
- the previous literature on the use of information theory in statistical inference justifies its use by arguments of information transmission, a set of postulates supposed to be obvious, or by analogy with entropy in statistical mechanics none of which are convincing from the point of view of statistical inference (Kendall (1973), Hart (1971)).

Thus the results of Larimore (1983a) give a solid theoretical justification for the use of the negative entropy measure in a general setting which makes possible the further general development of predictive inference statistical methods.

For the parametric case, the very general considerations above simplify somewhat. For the structure determination problems, an estimator of the form as in (2-7) associates a parameter estimate $\hat{\theta}(x)$ with each possible value x of the sample space. To simplify the discussion in this section, we can consider that the informative experiment x (fit set) and predictive experiment y (check set) are independently and identically distributed K -dimensional vectors. In Section 4, the general dependent time series analysis case will be discussed. We predict the density of y by $p(y, \hat{\theta}(x))$ where $p(y, \theta)$ is the parameterized class of densities for y . The negative entropy measure (2-6) reduces in the parametric case to that suggested by Akaike (1973) and is expressible as

$$R(p, \hat{\theta}) = E_x K(p, \hat{\theta}) = E_x \int p(y, \theta_x) \log \frac{p(y, \theta_x)}{p(y, \hat{\theta}(x))} dy \quad (2-8)$$

where θ denotes the true value of the parameter θ and where E_x denotes expectation with respect to the informative sample x . That the estimator $\hat{\theta}(x)$ may involve different model orders or structures as in (2-7) is no conceptual difficulty although it may complicate the evaluation of the negentropy (2-8). The predictive sample y (check set) is never actually drawn, but we wish to devise decision procedures which would optimally predict in terms of (2-8).

The major statistical problem is to devise model-estimation/structure-determination schemes which come close to minimizing the negentropy. A major step in that direction was made by Akaike (1973) in proposing an extension of the maximum likelihood method to compare different model orders or structures. Suppose that $\hat{\theta}_k(x)$ is the maximum likelihood estimator for a given restriction of the parameters θ to a subspace H_k that is defined for every x in the sample space. Then we wish to partition the sample space into the disjoint subsets X_k so that for $x \in X_k$, the estimator

$$\hat{\theta}_{\hat{k}(x)} = \hat{\theta}_k(x) \text{ for } x \in X_k \quad (2-9)$$

is used. Akaike (1973) shows that asymptotically for nested models, an unbiased estimate of the negentropy using the maximum likelihood model $\hat{\theta}_k(x)$ for the whole sample space $x \in X$ is given by the Akaike information criterion (AIC) defined by

$$AIC(k) = -2 \ln p(x, \hat{\theta}_k(x)) + 2K(k) \quad (2-10)$$

where $K(k)$ is the dimension of H_k , i.e., the number of parameters estimated. The Minimum AIC Estimate (MAICE) proposed by Akaike (1973, 1974b) is to partition the sample space so that X_k is the set of sample points for which

$$AIC(k) < AIC(j) \text{ for } j \neq k \quad (2-11)$$

Then the MAICE estimate is

$$\hat{\theta}_{MAICE}(x) = \hat{\theta}_k(x) \text{ for } x \in X_k \quad (2-12)$$

so that on the set X_k , $\hat{\theta}_{MAICE}(x)$ is the maximum likelihood estimate $\hat{\theta}_k(x)$ corresponding to the model structure k with minimum AIC.

For autoregressive models, Shibata (1981a) has studied the MAICE and other asymptotically equivalent procedures for model-estimation/order-determination. He adopted a spectral measure of accuracy that is asymptotically equivalent to the negentropy. He showed that asymptotically for large sample, MAICE minimizes the negentropy measure of accuracy (2-8), which will be called *entropy efficiency*. Hence MAICE is asymptotically an optimal procedure for choosing autoregressive models. Shibata (1981b) also shows MAICE as asymptotically optimal for regression problems which involve nonnested multiple comparisons. Other procedures for model order determination have been proposed (Bhansali and Downham, 1977; Schwarz, 1978) which emphasize the choice of true model order asymptotically for large samples, which is called *order consistency*. In most real problems the true order is infinite, and even if such a fiction were to exist, a predictive criterion is much more intuitive in most applications. Shibata (1983) has shown

that order consistency and entropy efficiency are mutually exclusive so that a choice is required as to which of these criteria is most important. In particular, Shabata has shown that an order consistent procedure cannot be entropy efficient, and that an entropy efficient procedure will not be order consistent.

Turing now to a different general problem, that of exact small sample inference, predictive inference and entropy provides a new approach to the problem. Past approaches to the small sample inference problem have involved a number of *ad hoc* procedures. The entropy measure provides a sound fundamental measure of the approximation error in predicting the density of the future experiment. One of the past approaches has involved the *estimative method* where the predictive density is restricted to lie in the class of densities assumed to contain the true. Recent results have shown that the use of more general predictive densities can give more optimal results as measured in terms of the entropy measure of model approximation error (Murray, 1979, 1977; Aitchison, 1975). The optimal predictive density has been derived in the class of invariant densities minimizing the entropy measure. This was derived before the justification of the entropy measure based upon the sufficiency principle. As shown in Chapter 6, this more general and optimal predictive density can be considerably better as measured by the entropy.

In time series analysis, the advantage of the approach using predictive inference and entropy is that it provides a sound theoretical framework in terms of model approximation for the direct comparison of very general time series analysis models including:

- Consideration of many complex hypotheses
- Comparison of nonnested hypotheses
- Comparison of dynamic models of different dynamic (state) orders
- Consideration of models fitted over different data sets for detecting abrupt changes
- Consideration of different adaptation rates for doing optimal model tracking

The comparison of such diverse models is inherent in adaptive time series analysis and abrupt change detection, and previous investigations have not had available such a sound and general framework for solving these difficult problems.

3. CONSTRAINED NONNESTED MULTIPLE COMPARISON OF MODELS

In this chapter, the general problem of nonnested multiple comparison is considered. In order to develop a general theory, consideration is restricted to comparing models that are the result of constrained maximum likelihood estimation. The objective of the discussion is to generalize the currently available procedures for nonnested multiple comparison in the constrained maximum likelihood context.

The approach is to view the fitting of each alternative parametric model form as an approximation procedure, which includes the notion that the true model is in general not contained in the class of parametric models considered in the model fitting. This is a departure from previous approaches that involve primarily asymptotic arguments where the parametric models approach the true model as the sample size becomes large. Such an asymptotic argument begs the question of model approximation since asymptotically there is no error in the approximation. It is very important in practice to determine the extent to which the asymptotic approximations are accurate in moderate or small samples.

Another area of weakness in available approaches is the assumption of nesting in comparing models using entropy methods. The derivations of Akaike involve the assumption of nested models which considerably restricts the applications of the methods. In practice, the AIC criterion has been applied in a much wider context than the comparison of nested models.

The results of this chapter will provide the basis of much more general decision procedures for adaptive time series analysis. In these problems, the comparison of different models based upon different intervals of data are compared to determine if an abrupt change has occurred. Previous entropy methods have only compared different models for a given interval of data.

3.1 Constrained Maximum Likelihood Estimation

The first result to be discussed is the generalization of the usual maximum likelihood theory to the constrained case. The regular case is considered where the log likelihood function is expandable in a Taylor series (Cox and Hinkley, 1974, p. 281). These conditions permit the interchange of expectation and differentiation.

Following the notation of Larimore (1986d, contained in Appendix A), let $l(x, \theta)$ denote the log likelihood function of the informative sample x considered as a function of the parameters θ . Denote by E the expectation with respect to the true density $p(x, \bar{\theta})$ with true parameter $\bar{\theta}$. The negative entropy measure is used as the measure of approximation to the true density by an

approximating density $p(x, \theta^k)$ with parameter value θ^k from the subspace Θ_k of parameters. The projection $\tilde{\theta}^k$ of $\tilde{\theta}$ onto the subspace Θ_k is defined as the parameter value θ^k minimizing

$$R_x(\tilde{\theta}, \theta^k) = El(x, \tilde{\theta}) - El(x, \theta^k) \quad (3-1)$$

so that the projection $\tilde{\theta}^k$ satisfies the condition

$$El'(x, \tilde{\theta}^k) = 0, \quad (3-2)$$

where $'$ denotes the derivative with respect to θ^k . The minimum is unique if and only if the Hessian D_x^k given by $D_x^k = El''(x, \tilde{\theta}^k)$ is positive definite. Thus for a constrained class of models, the projection of the true parameter value defines the best approximation to the true density in the class of approximating densities.

Consider now the constrained maximum likelihood estimate $\hat{\theta}^k$ in the subspace of parameters Θ_k satisfying the likelihood equation

$$l'(x, \hat{\theta}^k) = 0 \quad (3-3)$$

Then under the regularity conditions, we have for a positive definite Hessian D_x^k and asymptotically for large informative sample x that

- $\hat{\theta}^k$ is an unbiased estimator of $\tilde{\theta}^k$
- the estimation error covariance matrix is

$$E(\hat{\theta}^k - \tilde{\theta}^k)(\hat{\theta}^k - \tilde{\theta}^k)^T = (D_x^k)^{-1} E(l'^T(x, \tilde{\theta}^k) l'(x, \tilde{\theta}^k)) (D_x^k)^{-1} \quad (3-4)$$

For the unconstrained case, the middle term is the Fisher information matrix and is equal to minus the Hessian D_x^k , but in the general constrained case this is not true.

Now consider the likelihood $l(y|x, \theta)$ of the predictive experiment y conditioned on the informative experiment x . From the above results, the negentropy can be easily determined. Expanding the log likelihood to second order and taking expectation gives the negative entropy as

$$R_{y|x}(\hat{\theta}, \tilde{\theta}^k) = -\frac{1}{2} \|\hat{\theta} - \tilde{\theta}^k\|_{D_x^k}^2 + R_{y|x}(\tilde{\theta}, \tilde{\theta}^k) \quad (3-5)$$

which holds asymptotically for large informative sample. Note that the second term is exact with no approximation in small samples. The first term is an approximation involving the variation of the maximum likelihood estimate locally about the projection $\tilde{\theta}^k$. Thus the bias part of the error in constraining the model is captured exactly.

3.2 Unbiased Estimation of Entropy

In the previous discussion of entropy, the measure is considered as a measure of approximation error between the true and approximating density. In practice, the true density is unknown and it is desired to obtain an estimate of the negative entropy based upon the observed informative sample. To simplify the discussion, the case of x and y independent is considered. An accurate estimate of the negative entropy was first obtained by Akaike using the log likelihood as an estimate of the entropy with a correction for the bias. The Akaike information criterion (AIC)

$$AIC(k) = -2\log p(x, \hat{\theta}^k(x)) + 2K(k) \quad (3-6)$$

was derived as an unbiased estimate of the entropy where $K(k)$ is the number of parameters adjusted in fitting the maximum likelihood estimates. The second term adjusts for the bias in estimating the entropy using the informative sample and adjusting the parameters in fitting. Akaike (1973) originally derived the AIC as an unbiased estimate for the relative comparison of the prediction error in comparing two nested models. The nesting is also important in that derivation because the models are not only nested but asymptotically approach the true model.

In the more general case of constrained maximum likelihood estimation, a difficulty occurs in the estimation of the negentropy. Consider as above the case of x and y independent and identically distributed. As derived in Appendix A, the expected log likelihood difference of the informative sample is

$$E[l(x, \tilde{\theta}) - l(x, \hat{\theta}^k)] = -tr(D_x^k)^{-1} E\{l'^T(x, \hat{\theta}^k) l'(x, \hat{\theta}^k)\} + R_y(\tilde{\theta}, \hat{\theta}^k) \quad (3-7)$$

In the unconstrained case, asymptotically for large informative sample the trace term is equal to the number of parameters estimated. Unfortunately in the general constrained case, the Hessian is not equal to the Fisher information matrix. The trace then depends upon the expectation with respect to the true unknown density of the first and second derivatives of the log likelihood function at the projection $\tilde{\theta}^k$. This cannot be computed in general since the true parameter is unknown.

In cases where the Hessian and Fisher information are equal so that the trace is equal to the number of parameters, then two different parametric model structures, say θ^k and θ^j can be compared using (3-7) as

$$E[l(x, \hat{\theta}^k) - l(x, \hat{\theta}^j)] = -dim(\theta^k) - dim(\theta^j) + R(\tilde{\theta}, \hat{\theta}^j) - R(\tilde{\theta}, \hat{\theta}^k) \quad (3-8)$$

which is equivalent to the AIC. The derivation for constrained maximum likelihood estimates makes clear some of the assumptions in previous entropy methods. The major difficulty is caused by the variation of the Fisher information or Hessian as the parameter values change. In the

derivation above, the issue of nesting does not arise.

3.3 Nested Tests

In the case of nested tests, the MAICE criterion reduces to the usual generalized likelihood ratio (GLR) test. The threshold and resulting probability of rejecting the null hypothesis, i.e. the size of the test, depends upon the number of additional parameters in the more general model.

In comparing two hypotheses H_0 and H_1 , the AIC criterion is to choose according to the sign of the quantity

$$AIC(H_0) - AIC(H_1) = -2 \log \frac{p(x, \hat{\theta}^0)}{p(x, \hat{\theta}^1)} + 2[K(0) - K(1)] \quad (3-9)$$

The AIC criterion in the nested case is equivalent to the decision rule

$$\begin{aligned} \text{choose } H_0 & \text{ if } -2 \log \lambda < 2[K(1) - K(0)] \\ \text{choose } H_1 & \text{ if } -2 \log \lambda \geq 2[K(1) - K(0)] \end{aligned} \quad (3-10)$$

where the *generalized likelihood ratio* λ is defined by

$$\lambda = \frac{p(x, \hat{\theta}^1)}{p(x, \hat{\theta}^0)} \quad (3-11)$$

The threshold $2[K(1) - K(0)]$ is precisely twice the number of additional parameters under the hypothesis H_1 .

In the case of a normal class of densities, the size α of the test is easily determined since the GLR statistic λ is chi-squared on $K(1) - K(0)$ degrees of freedom under the null hypothesis H_0 . Thus α is given as the solution of the relationship

$$X_{\alpha, n}^2 = 2n \quad (3-12)$$

where n is the number of additional parameters and $X_{\alpha, n}^2$ is the α probability point of the cumulative chi-squared distribution on n degrees of freedom. Solving for α as a function of n gives the size of the test as a function of the number n of additional parameters as shown in Table 1. Note that the traditional α levels of 0.10, 0.05, 0.01, 0.005 and 0.001 correspond respectively to about 4, 8, 16, 20, and 30 additional parameters. It has been known for a long time that in composite tests or repeated applications of simple additions of a single variable that the significance level is considerably reduced such as in step-wise regression. The entropy approach makes explicit the change in the size α of the test with the number of additional estimated parameters.

Number n of Additional Parameters	1	2	3	4	5	8	11	16	20	30
Probability α of Rejecting Null Hypothesis	0.167	0.144	0.115	0.094	0.080	0.043	0.024	0.010	0.005	0.0008

Table 1. Dependence of Significance Level α on the Number n of Additional Parameters

4. TIME SERIES ANALYSIS USING ENTROPY METHODS

Methods of predictive inference and entropy offer a number of advantages in the analysis of time series not available in other methods. In this chapter the basic time series analysis methods are described, while in the following chapter adaptive methods for time series analysis are developed using predictive inference and entropy. First the topic of the achievable accuracy of spectral analysis is addressed by relating the entropy measure directly to a relative squared error in estimating the power spectrum. Following this is a discussion of the Markovian representation of time series in terms of state space models which will be very useful in representing time varying models of time series. The canonical variate analysis approach to time series is then described which forms the basis of the adaptive time series analysis methods developed in the following chapter.

4.1 Achievable Spectral Accuracy

In this section, the informative and predictive samples will be denoted by u and v respectively to allow for the traditional use of x and y for random processes. Consider the problem of identifying a model for a pair (x, y_t) of multiple stationary time series where x_t and y_t are exogenous and endogenous time series respectively. Consider linear stochastic models in the form of a linear difference equation

$$y_t = q_t + \sum_{\tau=0}^{\infty} h(t-\tau; \theta) x_{t-\tau} = q_t + r_t \quad (4-1)$$

where $h(t; \theta)$ is a causal linear system giving the response in y_t to the past exogenous inputs $x_{t-\tau}$, and where q_t is white noise. Suppose that the probability density of the process is parameterized by θ . The exogenous variable x_t will be considered as exactly observed, and the problem of modeling y_t conditional on x_t is considered so that prediction of y_t from x_t based upon such a model is the principle problem. This also includes the problem of no exogenous variable so that only y_t is observed.

We wish to investigate the achievable accuracy in estimating a model for the process. In particular, the entropy measure will be developed to obtain the relationship between the number of parameters estimated in the model and the relative squared error in estimating the power spectrum. An example illustrates the effect on the spectral estimation error due to the particular class of parametric models used in the identification and the number of parameters estimated.

The predictive inference setup of Chapter 2 is considered where the primary interest is in the asymptotic behavior with large sample size of both the informative and predictive samples. Consider an observed informative sample $u^T = (x_1^T, y_1^T, \dots, x_N^T, y_N^T)$ of size N used to estimate the process model, and similarly consider a conceptual predictive sample v of size M used to evaluate the accuracy of the estimated model. The predictive sample is assumed to be identically distributed but independent of the informative sample. Consider the problem of inference on the parametric class $\{p(v, \theta), \theta \in \Theta\}$ of models with probability densities $p(v, \theta)$ based upon the informative sample u . Consider the conceptual repeated sampling experiment where on each trial the samples u and v are each drawn independently from the process $S(\omega, \bar{\theta})$ with $\bar{\theta}$ assumed to be the true parameter value. An estimative model $\hat{p} = p(v, \hat{\theta}(u))$ is chosen for the density of v by some parameter estimation scheme $\hat{\theta}(u)$. For a stationary process, the negative entropy (2-6) is linear in the predictive sample size M , so it is more useful to consider the per sample negentropy. To this end, define the *per sample negentropy* denoted $I(p, \hat{p})$. As derived in Appendix B, the I-divergence is given asymptotically by

$$\begin{aligned} I(S, \hat{S}) &= \lim_{M \rightarrow \infty} \frac{1}{M} E_u \int p(v, \theta_u) \log \frac{p(v, \theta_u)}{p(v, \hat{\theta}(u))} dv \\ &= \frac{1}{4} E_u \int_{-\pi}^{\pi} \text{tr} \{ S_{\pi\pi}^{-1}(\omega) [\hat{S}_{\pi\pi}(\omega) - S_{\pi\pi}(\omega)]^2 \} \frac{d\omega}{2\pi} \\ &\quad + \frac{1}{2} E_u \int_{-\pi}^{\pi} \text{tr} \{ \hat{S}_{\pi\pi}^{-1} [H(\omega) - \hat{H}(\omega)] S_{xx}(\omega) [H(\omega) - \hat{H}(\omega)]^* \} \frac{d\omega}{2\pi} \end{aligned} \quad (4-2)$$

where E_u denotes expectation relative to the informative sample u .

In the multiple time series case, the spectral measure (4-2) has an intuitive interpretation in terms of principal components of the power spectrum in the frequency domain. Principle component representations of the spectral matrices $S_{\pi\pi}(\omega)$ and $S_{xx}(\omega)$ have the form

$$J(\omega) S_{\pi\pi}(\omega) J^*(\omega) = D(\omega) \quad , \quad L(\omega) S_{xx}(\omega) L^*(\omega) = E(\omega) \quad (4-3)$$

where $J(\omega)$ and $L(\omega)$ given as a function of frequency ω are unitary matrix transformations so $J(\omega)J^*(\omega) = I = L(\omega)L^*(\omega)$ which diagonalize $S_{xx}(\omega)$ and $S_{\pi\pi}(\omega)$ respectively and where $D(\omega)$ and $E(\omega)$ are diagonal matrices. Filtering $x(t)$ with transfer function $L(\omega)$ gives the principal component process $\bar{x}(t)$ which is expressed in the frequency domain as $\bar{X}(\omega) = L(\omega)X(\omega)$, and which has the diagonal spectral matrix $E(\omega)$, and similarly for $q(t)$.

The spectral measure (4-2) is shown in Appendix B to be

$$\begin{aligned}
I(S, \hat{S}) = & \frac{1}{4} \sum_i \int_{-\pi}^{\pi} \left[\frac{\hat{D}_{ii}(\omega) - D_{ii}(\omega)}{\hat{D}_{ii}(\omega)} \right]^2 \frac{d\omega}{4\pi} + \frac{1}{2} \sum_{i \neq j} \int_{-\pi}^{\pi} \frac{|\hat{D}_{ij}(\omega)|^2}{\hat{D}_{ii}(\omega)\hat{D}_{jj}(\omega)} \frac{d\omega}{4\pi} \\
& + \frac{1}{2} \sum_{i \neq j} \int_{-\pi}^{\pi} |\hat{G}_{ij}(\omega) - G_{ij}(\omega)|^2 \frac{D(\omega)_{ii}}{E(\omega)_{jj}} \frac{d\omega}{2\pi}
\end{aligned} \quad (4-4)$$

The first sum on the right hand side is the integrated squared relative error of the estimated co-spectra of the principal components, while the second term is the integrated squared coherency of the estimated spectrum $\hat{D}(\omega)$ which would be zero if $\hat{D}(\omega) = D(\omega)$. Thus the measure (4-2) has a clear interpretation in the multivariate case when the true spectrum $D(\omega)$ is diagonal but where the approximating spectrum $\hat{D}(\omega)$ is permitted arbitrary coherency among components. The third term in the spectral measure (4-4) is asymptotically equivalent to replacing $\hat{S}_{xx}(\omega)$ by $S_{xx}(\omega)$. This term is invariant to the unitary transformations $J(\omega)$ and $L(\omega)$ where $G(\omega) = J(\omega)H(\omega)L^*(\omega)$ is the transfer function $H(\omega)$ expressed in the coordinate frame of the principal component series $\bar{x}(t)$ and $\bar{y}(t)$. The squared magnitude error $|\hat{G}_{ij}(\omega) - G_{ij}(\omega)|^2$ in the i, j element of the transfer function is weighted by the input signal to output noise ratio $D(\omega)_{ii}/E(\omega)_{jj}$ for the pair (i, j) .

The spectral measure of accuracy can be bounded in terms of the number of parameters estimated. Suppose first for simplicity that the parametric class of models contains the true process and that k parameters are estimated. Then by Appendix B using the Cramer-Rao lower bound, the per sample negentropy is bounded by

$$E_N I(S, \hat{S}) = E_N \frac{1}{2N} (\hat{\theta} - \theta)^T F (\hat{\theta} - \theta) \geq \frac{1}{2N} \text{tr} F^{-1} F = \frac{k}{2N} \quad (4-5)$$

with equality achieved asymptotically for large informative sample N . This implies the bound on the achievable accuracy in spectral estimation given by

$$\begin{aligned}
\frac{k}{2N} \leq & \frac{1}{4} E_N \int_{-\pi}^{\pi} \text{tr} \{ S_{\eta\eta}^{-1}(\omega) [\hat{S}_{\eta\eta}(\omega) - S_{\eta\eta}(\omega)]^2 \} \frac{d\omega}{2\pi} \\
& + \frac{1}{2} E_N \int_{-\pi}^{\pi} \text{tr} \{ \hat{S}_{\eta\eta}^{-1} [H(\omega) - \hat{H}(\omega)] S_{xx}(\omega) [H(\omega) - \hat{H}(\omega)]^T \} \frac{d\omega}{2\pi}
\end{aligned} \quad (4-6)$$

In the more general case where the order is infinite and the MAICE procedure for choosing model order k is used, then Shibata (1983) has shown the following result. For each informative sample size N there is an optimal order $k^*(N)$ which minimizes the tradeoff between variability and bias in the entropy measure

$$R(\hat{\theta}, \hat{\theta}^k) = -\frac{1}{2} \|\hat{\theta}^k - \hat{\theta}\|_{D_{\hat{\theta}}^k}^2 + R(\hat{\theta}, \hat{\theta}^k) = \frac{k}{2N} + R(\hat{\theta}, \hat{\theta}^k) \quad (4-7)$$

Then asymptotically for large informative sample N , the negative entropy of the MAICE model selection procedure is exactly that of the model selection procedure using a fixed number of parameters equal to the optimal order k^* . Thus even in the case of using an entropy efficient model selection procedure where the true model order is infinite, the achievable accuracy of spectral estimation (4-6) can be bounded by the function (4-7) of the sample size N with $k = k^*$.

To illustrate the use of the lower bound on the achievable accuracy, consider the ARMA (4,3) process

$$y_t = 1.3136 y_{t-1} - 1.4401 y_{t-2} + 1.0919 y_{t-3} - 0.83527 y_{t-4} \\ + w_t + 0.17921 w_{t-1} + 0.82020 w_{t-2} + 0.26764 w_{t-3} \quad (4-8)$$

with the noise variance of w as $Q = 1.7258 \times 10^{-2}$. This process was analyzed by Gersch and Sharp (1973) and Akaike (1974b) to show the increased accuracy of ARMA models over AR models. For a sample size of 800, the optimal order was found to correspond to $k^* = 18$ for fitting an AR model to the data. Akaike (1974b) fitted several models to simulated data using AR, ARMA, and Hanning window methods. Figure 1 shows the variability term of the spectral error as a function of frequency for the various model fitting procedures. Since the optimal order was used for the AR model, the bias is also included. The ARMA model has no bias since the full order is chosen with high probability. On the other hand, the Hanning window has significant bias since a fixed bandwidth is used to spectrally smooth the data at all frequencies, and this bandwidth is not sufficient to estimate the sharp peak without bias. Use of a wider bandwidth would increase the already large error at all the other frequencies. The AR and ARMA methods are clearly adaptive in that the methods have a greater bandwidth to accommodate the rapid changes (large second derivative or curvature) near the peaks and troughs but lower bandwidth in regions with low curvature. The greater parametric efficiency of the ARMA method is clearly depicted in these regions of low curvature. Repeated simulation of the time series data from the ARMA(4,3) model and the maximum likelihood estimation of ARMA models with MAICE confirms that the lower bound indeed gives an accurate description of the spectral estimation error in practice (Larimore, Mahmood, and Mehra, 1984). Independent methods have been developed for obtaining simultaneous confidence bands on spectral estimates (Larimore, 1986c). Such confidence bands are proportional to the integrand of the spectral measure of accuracy so that the integrand gives an accurate measure of spectral error at each frequency as well.

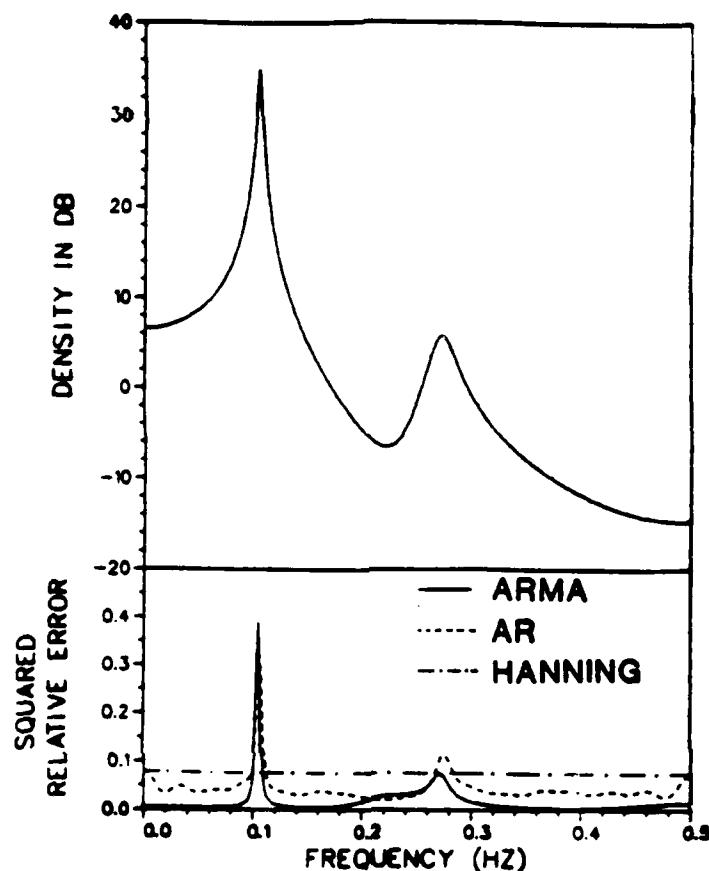


Figure 1. Spectral Accuracy of Different Model Fitting Procedures (lower curves) in Approximating the True Spectral Density (upper curve).

4.2 Markovian Models of Time Series

In this section Markovian or state space models of time series are reviewed. Such models have not been widely used in time series analysis, although there is wide spread use of such models in filtering and prediction with numerous applications in engineering. State space models have a number of advantages in time series analysis that are attractive for automatic implementation on microprocessors using the canonical variate analysis method discussed in the next section. Such procedures allow the automatic selection of model state order using entropy methods and lend themselves to adaptive methods for time varying processes discussed in the next Chapter.

The starting point of any approach is the joint probability distribution of the past and future observations $p(f, p, \theta)$ where p_t are the past inputs u_t and outputs y_t up to time t and f_t are the outputs y_t in the future at time t defined by

$$p_i^T = (\cdots y_{i-1}^T, u_{i-1}^T, y_i^T, u_i^T) \quad , \quad f_i^T = (y_{i+1}^T, y_{i+2}^T, \cdots) \quad (4-9)$$

and θ is a vector of parameters indexing the model. A fundamental property of a Markov process of finite state order is the existence of a finite dimensional state x_i which is a linear function

$$x_i = Cp_i \quad (4-10)$$

of the past p_i . The state x_i has the property that the distribution of the future f_i conditioned on the past p_i is identical to that of the future f_i conditioned on the finite dimensional state x_i , so

$$p(f_i | p_i, \theta) = p(f_i | x_i, \theta) \quad (4-11)$$

Thus, only a finite amount of information from the past is relevant to the future evolution of the process.

A stationary Markov process of some particular state order can be represented by a vector difference equation with the general form (Lindquist and Pavon, 1981)

$$x_{i+1} = \Phi x_i + Gu_i + w_i \quad (4-12)$$

$$y_i = Hx_i + Au_i + Bw_i + v_i \quad (4-13)$$

where u is an input vector process, y is the output vector, x is the state vector, and w and v are white noise processes that are independent with covariance matrices Q and R respectively. The matrices Φ , A , B , G , and H determine the dynamics of the process and correlational characteristics of the disturbances. The various matrices are considered as functions of the parameters θ specifying the process. The white noise processes model the covariance structure of the error in predicting y from u .

For time series analysis and system identification, the parameterization of the model is an important issue. The elements of all of the matrices of the state space model (4-12) and (4-13) and noise covariances are not independent parameters of the model. In fact for each distinct probability distribution there is an equivalence class of models of the form (4-12) and (4-13) with the same distribution. It can be shown (Candy, Bullock, and Warren, 1979) that the number of independent parameters is

$$K(k) = 2kn + n(n+1) + 2km + nm \quad (4-14)$$

where k , n , and m are the vector dimensions of the state x_i , outputs y_i , and inputs u_i respectively. If there is no instantaneous feedforward so $A=0$, then the term nm is deleted, while if there is no input so $A=G=0$ the terms $km+nm$ are deleted.

The state space parameterization (4-12) and (4-13) is not unique, however in the next section a well conditioned procedure for selecting a unique model for the equivalence class will be described. For an individual state space model there exists a corresponding ARMA model and visa versa. However the two classes are not equivalent as classes. In general there is no ARMA class of models equivalent of a particular state space class. The ARMA class has one major difficulty - there is no global parameterization of the state space models of a given order. The difficulty is in the ARMA representation which becomes singular at certain models such as one involving the cancellation of a pole and a zero. This causes great difficulty in numerical methods in attempting to automatically identify higher order models which may involve such cancellations of poles and zeros.

The major advantage of the state space models is the availability of efficient and numerically well conditioned procedures for model identification discussed in the next section, and the explicit Markov structure allows for the the development of direct adaptation procedures developed in the next chapter.

4.3 Canonical Variate Analysis of Time Series

The canonical variate analysis method for identification of state space time series models is described in this section. The methods for the determination of the state order and selection of the state using concepts of canonical variate analysis are first discussed. The determination of the state space model is then computed by simple regression. The computation involves primarily a singular value decomposition of the sample covariance matrix of the process.

A generalization of the canonical variate analysis method has recently provided a completely general solution to the static reduced rank stochastic prediction problem which is well defined statistically and computationally even when some or all of the various covariance matrices are singular (Larimore, 1986b). All other previous methods in the statistical literature do not address the general problem. This result is the foundation of the time series analysis methods using predictive inference and entropy including the adaptive time series methods.

The original development of the canonical correlation analysis method of mathematical statistics was by Hotelling (1936; see also Anderson, 1958). The application of canonical variate analysis to stochastic realization theory and system identification was done in the pioneering work of Akaike (1974a, 1975, 1976). This initial work has a number of limitations such as no system inputs, no additive measurement noise, substantial computational burden involving numerous SVD's, a heuristic set of decisions for choosing a basis for representation of the system, and a number of approximations including computation of the AIC criterion for decision on model

order.

Some important generalizations and improvements in Akaike's canonical correlation method have recently been made by Larimore (1983b). These include generalization to systems with additive measurement noise and with inputs including feedback controls. A major departure of the approach from previous work is the use of a single canonical variate analysis to optimally choose k linear combinations of the past for prediction of the future. The very natural measure of quadratically weighted prediction errors at possibly all future time steps is used. Formulated as such a prediction problem, it is shown how a generalized canonical variate analysis gives the solution explicitly. The interpretation of canonical variates as optimal predictors is central in motivating interest in such a problem formulation and is scarcely found in the statistical literature (Larimore, 1986b). The optimal k -order predictors are not in general recursively computable, but the optimal state-space structure for approximating them is expressed simply in terms of the canonical variate analysis. The problem of finding an optimal Hankel norm reduced order model (Adamjan et al, 1971; Kung and Lin, 1981) is related to the canonical variate approach (Camuto and Menga, 1982; Larimore, 1983b). The balanced realization method is a particular case of the generalized canonical variate analysis (Desai and Pal, 1984).

To more concisely discuss the canonical variate method, the results in Larimore (1983b, 1986b) are briefly reviewed. Consider the problem of choosing an optimal system or model of specified order for use in predicting the future evolution of the process. As in Section 4.2, consider the past p_i of the inputs u_i and outputs y_i before time i and the future of the outputs y_i at time i or later so

$$p_i^T = (\cdots \mathcal{O}_{i-1}^T \mu_{i-1}^T \mathcal{O}_{i-1}^T \mu_i^T) \quad , \quad f_i^T = (y_{i+1}^T \mathcal{O}_{i+2}^T, \cdots) \quad (4-15)$$

We assume that the processes u_i and y_i are jointly stationary and denote the covariance matrices among f_i and p_i as Σ_{ff} , Σ_{pp} , and Σ_{fp} .

The major interest is in determining a specified number k of linear combinations of the past p_i which allow optimal prediction of the future f_i . The set of k linear combinations of the past p_i are denoted as a $k \times 1$ vector m_i and are considered as k -order memory of the past. The optimal linear prediction \hat{f}_i of the future f_i , which is a function of a reduced order memory m_i , is measured in terms of the prediction error

$$E\{\|f_i - \hat{f}_i\|_{\Sigma_{ff}}^2\} = E\{(f_i - \hat{f}_i)^T \Sigma_{ff}^\dagger (f_i - \hat{f}_i)\} \quad (4-16)$$

where E is the expectation operation and \dagger denotes the pseudoinverse of a matrix. The optimal prediction problem is to determine an optimal k -order memory

$$m_i = J_k p_i \quad (4-17)$$

by choosing the k rows of J_k such that the optimal linear predictor $\hat{f}_i(m_i)$ based on m_i minimizes the prediction error.

As derived in Larimore (1986b), the solution to this problem in the completely general case where the matrices Σ_{ff} , Σ_{pp} , and Σ_{fp} may be singular is given by the generalized singular value decomposition as stated in the following theorem.

Theorem 1. Consider the problem of choosing k linear combinations $m_i = J_k p_i$ of p_i for predicting f_i such that (4-16) is minimized where Σ_{pp} and Σ_{ff} are possibly singular positive semidefinite symmetric matrices with ranks \bar{m} and \bar{n} respectively. Then the existence and uniqueness of solutions are completely characterized by the $(\Sigma_{pp}, \Sigma_{ff})$ -generalized singular value decomposition which guarantees the existence of matrices J , L , and generalized singular values $\gamma_1, \dots, \gamma_r$ such that

$$J \Sigma_{pp} J^T = I_{\bar{m}}, \quad L \Sigma_{ff} L^T = I_{\bar{n}}, \quad J \Sigma_{fp} L^T = \text{Diag}(\gamma_1 \geq \dots \geq \gamma_r, 0, \dots, 0) \quad (4-18)$$

The solution is given by choosing the rows of J_k as the first k rows of J if the k -th singular value satisfies $\gamma_k > \gamma_{k+1}$. If there are r repeated singular values equal to γ_k , then there is an arbitrary selection from among the corresponding singular vectors, i.e. rows of J . The minimum value is

$$\min_{\text{rank}(J_k \Sigma_{pp} J_k^T) = k} \|f_i - \hat{f}_i\|_{\Sigma_{ff}}^2 = (1 - \gamma_1^2) + \dots + (1 - \gamma_k^2) \quad (4-19)$$

This result not only gives a complete characterization of the solutions in selecting optimal predictors m_k from the past p_i for prediction of the future f_i , but the reduction in prediction error for all possible selections of order k is given simply in terms of the generalized singular values. This is of great importance since it avoids having to do a considerable amount of computation to determine what selection of order is appropriate in a given problem.

The generalized CVA method allows the determination of the fit of the various state space models and the selection of the best model state order before computation of the state space models. Consider the general case of identifying a state space model: given the past of the related random process u_i and y_i , we wish to model and predict the future of y_i by a k -order state-space structure of the form (4-12) and (4-13)

$$x_{i+1} = \Phi x_i + G u_i + w_i \quad (4-20)$$

$$y_i = H x_i + A u_i + B w_i + v_i \quad (4-21)$$

In the computational problem given finite data, the past and future of the process are taken to be finite of length d lags so

$$p_i^T = (y_{i-1}^T, \dots, y_{i-d}^T, \mu_{i-1}^T, \dots, \mu_{i-d}^T) \quad , \quad f_i^T = (y_i^T, \dots, y_{i+d}^T) \quad (4-22)$$

Akaike (1976) proposed choosing the number d of lags by least squares autoregressive modeling using recursive least squares algorithms and choosing the number of lags as that minimizing the AIC criterion discussed below. This insures that a sufficient number of lags are used to capture all of the statistically significant behavior in the data. This procedure is easily generalized to include the case with inputs u_i . The generalized SVD of Theorem 1 determines a transformation J of the past that puts the state in a canonical form so that the memory $m_i = Jp_i$ contains the states ordered in terms of their importance in modeling the process. The optimal memory for a given order k then corresponds to selection of the first k states.

In order to decide on the model order to select, the Akaike information criterion (2-10) is computed where the number of parameters is determined from (4-16). Once the optimal k -order memory m_i is determined, state-space equations of the form (4-12) and (4-13) for approximating the process evolution are easily computed by a simple multiple regression procedure (Larimore, 1983b).

Since the CVA system identification procedure involves the state space model form, it has the major advantage that the model is globally identifiable so that the method is statistically well conditioned in contrast to ARMA modeling methods (Gevers and Wertz, 1982). Furthermore, since the computations are primarily a SVD, they are numerically stable and accurate with an upper bound on the required computations (Golub, 1969). Thus the method is completely reliable. It has been demonstrated as such in the time series analysis software Forecast Master that is commercially sold by SSI. From the theory of the CVA method (Larimore, Mahmood and Mehra, 1984), it can be shown that there are no difficulties such as biased estimates caused by the presence of a correlated feedback signal. The CVA method was demonstrated in real time identification and adaptive control of unstable aeroelastic wing flutter on a scale model F-16 aircraft in the NASA Langley Transonic Dynamics Wind Tunnel in February 1986.

5. ADAPTIVE TIME SERIES ANALYSIS

The state space model identification methods are developed in this chapter for adaptive time series analysis. The concepts of a changing Markov process are first discussed along with concepts of a piecewise constant model of the process that is constant over intervals of time. The approach to adaptation to slow changes using predictive inference and entropy is described. This leads to a model fitting criterion for choosing an optimal data interval that balances the decreasing sampling variability with increasing sample size against the increasing missmodeling error due to use of a constant model over an interval of data. In fitting models involving abrupt changes, the models fitted over various intervals are compared to determine if an abrupt change has occurred. This involves the comparison of models determined from data on different data intervals in predicting the error on a different interval. Several examples are given in using the procedure on changes involving the dynamics, noise excitation, measurement noise, and other changes.

Concepts of adaptive systems have been around since the 1950's involving various senses of adaptation. The present literature on the subject includes a number of methods such as recursive computational schemes, exponential forgetting, lattice computational methods, etc., which have certain "knobs" that allow tuning of the algorithm to accommodate changes in the characteristics of the actual processes. Reviews of these and related methods are contained in several recent special issues of technical journals and books (Special Issue on Adaptive Control, *Automatica*, Vol. 20, No. 5, 1985; Special Issue on Linear Adaptive Filtering, *IEEE Trans. on Information Theory*, Vol. 30, No. 2, 1984; Honig and Messerschmitt, 1984). While these methods do permit some degree of adaptation to process changes, the methods of adaptation are ad hoc, and no sound underlying statistical principle for adaptation is proposed or demonstrated. As might be expected, these methods can work poorly on certain cases because of the lack of a sound statistical basis.

In particular, the recursive prediction error and lattice methods are convenient due to their recursive form and provide an estimate at every observation (Friedlander, 1982a, 1982b, 1983; Ljung and Soderstrom, 1983). Also, the recursive algorithms can be used for adaptation by exponential weighting of the past data (Wellstead and Sanoff, 1981; Irving, 1979; Evans and Betz, 1982). But the rationale for exponential weighting has not been given a sound fundamental justification, but is used largely due to its ease of use. The choice of the exponential weight has been ad hoc and susceptible to misinterpretation of changing noise variance levels as time varying changes in the dynamics (Hagglund, 1983).

Adaptation to abrupt changes has been largely discussed in the fault detection literature. A comprehensive survey of fault detection methods is given by Willsky (1976). See also Mehra and Peachon (1971), Willsky and Jones (1974), Willsky (1980), and Isermann (1984). These methods have a number of shortcomings in detecting changes in dynamics, computational illconditioning, and excessive computational burden.

The central computation of any adaptive algorithm involves the extension of methods for identification of stationary time series. There are several difficulties with currently available methods and software for the identification of system dynamics and noise characteristics. Current methods include the self tuning regulator (STR) (Ljung, 1983; Astrom, 1973; Astrom et al, 1973, 1977), maximum likelihood estimation (MLE) (Mehra and Tyler, 1973; Larimore, 1981a), the Box-Jenkins (BJ) method (Box and Jenkins, 1976), and a variety of heuristic approaches. The current state of the art in both MLE and BJ require that an analyst be involved in the procedure, and the required number of computational iterations is not bounded. The STR has been applied successfully to simple processes, but is not completely reliable for general processes particularly when multi-input, multi-output systems are involved. In addition, the recursive prediction error algorithm used in the STR requires a good initial estimate and so is not suitable for short data where no apriori data is available. The heuristic approaches tend to be for special purposes and are rather unreliable in general applications.

5.1 Models for Changing Processes

The problem of modeling changing processes involves primarily two types of changes

- changes that are slow compared to the data interval used for identification
- abrupt changes occurring infrequently compared to the data interval used for identification.

If the changing process changes too rapidly or the abrupt changes occur too frequently relative to the data interval required for sufficiently accurate identification, then it is not possible to separate the actual system changes from the variability due to sampling.

Consider a time varying Markov process where the conditional probability of the future given the past depends upon time t so

$$p(f_t | p_t, \theta_t) = p(f_t | x_t, \theta_t) \quad (5-1)$$

where the state, defined as linear combinations of the past p_t , varies with time as

$$x_t = C_t p_t \quad (5-2)$$

In order to express the time evolution of the state of a Markov process in terms of a system of state space equations of the form

$$x_{i+1} = \Phi_i x_i + G_i u_i + w_i \quad (5-3)$$

$$y_i = H_i x_i + A_i u_i + B_i w_i + v_i \quad (5-4)$$

where the various matrices are time varying, it is necessary and sufficient that the conditional distribution of the future f_i , conditioned on the state x_i , have the form

$$p(f_i | x_i, \theta_i) = p(f_i | x_{i-1}, y_i, u_i, \theta_i) \quad (5-5)$$

This condition is essentially that the information in the state x_i for predicting the future f_i , is contained in the state x_{i-1} delayed by one time step and the present inputs u_i and outputs y_i . This condition is satisfied by most physical systems since the memory is stored as energy in physically describable states. If the system changes abruptly, there may also be an abrupt change in the input or a large noise innovation v_i associated with a significant change in the state of the system. Formulated in this way, it is apparent how the state space modeling methods are particularly useful.

In any modeling method based upon a finite sample of data, only a finite number of parameters can be determined which are much fewer in number than the number of data. Of the various possible methods for modeling, the simplest and least presumptive is the piecewise constant model which is constant over various intervals of data. Thus consider the model of the form (5-3) and (5-4) with piecewise constant parameters θ_i ,

$$x_{i+1} = \Phi_i x_i + G_i u_i + w_i \quad (5-6)$$

$$y_i = H_i x_i + A_i u_i + B_i w_i + v_i \quad (5-7)$$

The coefficient matrices Φ_i , G_i , H_i , A_i , and B_i are functions of the parameters θ_i which are constant over an interval of time T_i and change from one time interval to another.

In the following sections, adaptive time series analysis methods are developed by considering various hypotheses concerning slow and abrupt changes. The predictive inference and entropy methods provide a means of objectively comparing the vast multitude of such hypotheses entailed in the adaptive time series analysis problem.

5.2 Adaptation to Slow Variations

The problem of adaptation to slow variations is primarily that of determining the length of time interval to use in the time varying model (5-6) and (5-7). Consider the division of a section of data into 2^h subintervals of length 2^l samples where h and l are integers. Then the various hypotheses can be considered such as H_l : divide the interval into subintervals of length 2^l . For each subinterval I_j , for $j=1,2,\dots,2^h$, suppose a state space model denoted M_j is fitted using the CVA method with AIC used to select the best model state order.

By successive application of the Markov property, the joint probability density of the observations conditioned on the initial state is given by

$$\log p(Y_1, \dots, Y_{2^h} | Y_0, \theta_h) = \sum_{j=1}^{2^h} \log p(Y_j | Y_{j-1}, \theta_j) \quad (5-8)$$

where $\theta_h^T = (\theta_1^T, \dots, \theta_{2^h}^T)$ is the parameter vector for the composite model consisting of all of the models over the 2^h subintervals. This gives the log likelihood as the sum of the conditional log likelihoods on each subinterval.

Consider the entropy measure of the composite model θ_h . Using the asymptotic approximation (3-5), the negentropy is

$$R(\hat{\theta}, \hat{\theta}^h) = \frac{K_h}{2} + R(\hat{\theta}, \hat{\theta}^h) = \sum_{j=1}^{2^h} \left[\frac{K_j}{2} + R(\hat{\theta}, \hat{\theta}^j) \right] \quad (5-9)$$

where $\hat{\theta}^h$ and $\hat{\theta}^j$ are understood to denote the parameters constrained under the respective hypotheses involving estimation of these models. Note that there is a tradeoff between the first term which increases with finer subdivisions of the data and the second term which decreases as the number of parameters is increased with finer subdivisions of the data. A minimum of the negentropy defines the optimum subdivision of the data.

To estimate the negentropy from the sample, the AIC is used. From the definition of AIC, the AIC corresponding to (5-9) is given by

$$AIC(Y_1, \dots, Y_{2^h}, \hat{\theta}^h) = \sum_{j=1}^{2^h} AIC(Y_j, \hat{\theta}^j) \quad (5-10)$$

The optimal data length is chosen as the h minimizing the above AIC.

As an illustration of this scheme, it was applied to a problem in the 1978 Workshop on Spectral Estimation (Gerhardt, 1978) in estimating the instantaneous frequency of a sine wave with time varying frequency in the presence of interference and random noise. The data were

generated by the equation

$$y(t) = 1000 \cos(a(t)) + 100 \cos(b(t)) + n(t) \quad (5-11)$$

where the signal component has a time varying phase $a(t)$ and the interference component has phase $b(t)$ with the instantaneous frequencies given in Figure 2,

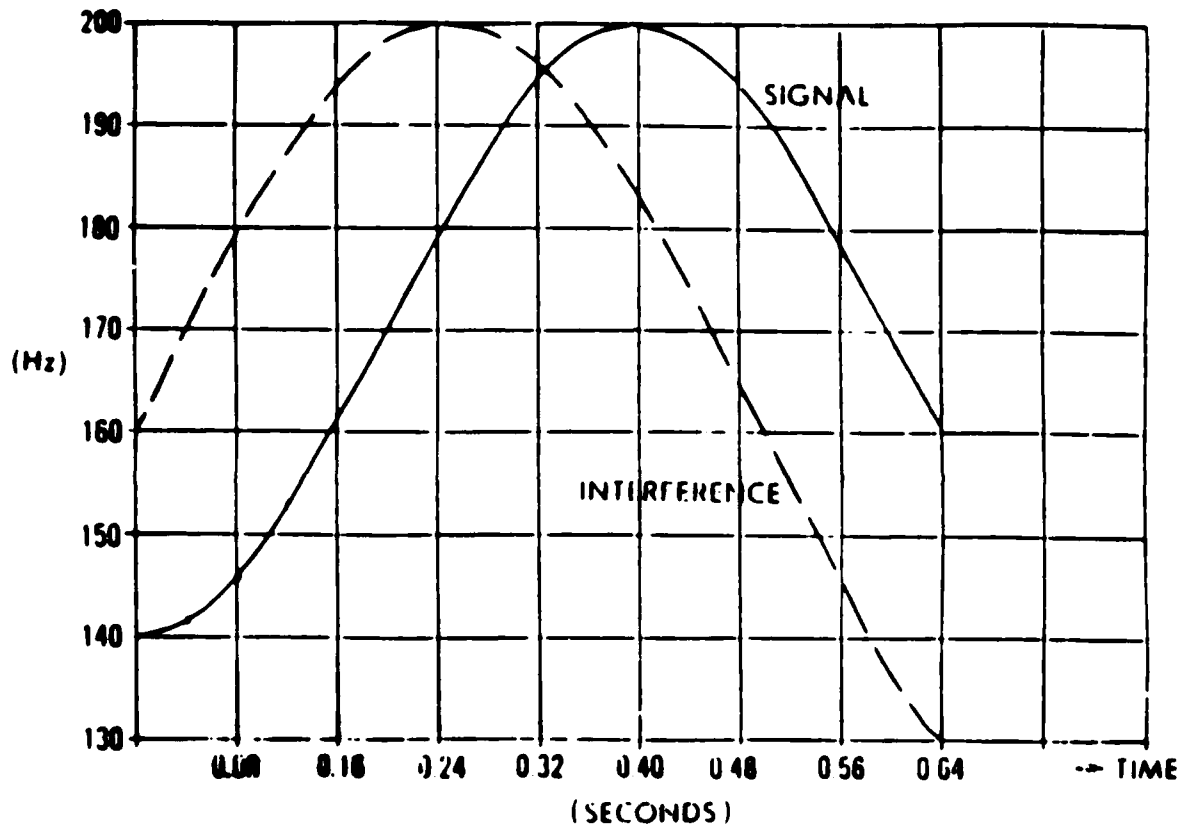


Figure 2. Instantaneous Frequency of Signal (solid) and Interference (dashed).

and where the noise is uniformly distributed with $-100 < n(i) < 100$. The participants were told to estimate the instantaneous frequency of the signal which was observed in the presence of interference and noise.

Table 2 gives the per sample AIC corresponding to the identified state space models for each of the subintervals used which were of lengths 16, 32, 64, and 128 samples. Also given are the per sample AIC's of the composite piecewise constant models for samples of 128. It is seen that the subinterval length producing the minimum average AIC for all of the data is 32 samples. This was then used as the optimal subinterval length for modeling the instantaneous frequency.

SAMPLE TIMES	NUMBER OF POINTS IN DATA SUBINTERVAL			
	16	32	64	128
6	12.40			
22	12.12	12.15		
38	12.39			
54	11.66	12.12	12.06	
70	12.37			
86	12.64	12.52		
102	12.53			
118	(12.28) 12.05	(12.17) 12.27	(12.16) 12.36	(12.36) 12.36
134	12.01			
150	12.84	12.26		
166	12.66			
182	11.52	12.03	12.29	
198	11.86			
214	11.98	12.02		
230	11.47			
246	(12.05) 12.06	(12.03) 11.79	(12.17) 12.06	(12.40) 12.40
262	12.03			
278	12.25	11.82		
294	11.39			
310	11.88	11.43	11.76	
326	12.08			
342	11.55	11.90		
358	12.08			
374	(11.74) 10.66	(11.70) 11.65	(11.78) 11.79	(11.75) 11.75
390				
ALL DATA	12.02	11.67	12.04	12.17

Table 2. Value of Per Sample AIC for Subintervals of the Sample, for the Average of 128 Points (), and for All Data.

The estimate of the instantaneous frequency was chosen as the maximum of the spectrum obtained from the CVA model fitted to the data. The estimated instantaneous frequency is given in Table 3 along with the other three best solutions obtained by the other participants in the workshop.

SAMPLE TIME	TRUE	WILEY AND CARMICHAEL	WEINER ET AL	ADAPTIVE CVA	MAXIMUM ENTROPY
32	141.47	141.47	141.69	142.13	141.7
64	145.73	145.73	145.88	145.13	146.2
96	152.37	152.37	152.69	152.22	152.6
128	160.73	160.73	160.71	161.85	160.3
160	170.00	170.00	170.22	169.30	169.9
192	179.27	179.27	179.48	178.40	179.9
224	187.63	187.63	187.72	188.06	188.2
256	194.27	194.27	194.33	193.90	194.8
288	198.53	198.53	197.83	197.22	198.0
320	200.00	200.00	199.72	200.94	201.5
352	198.53	198.53	197.89	198.35	198.3
384	194.27	194.27	193.93	194.66	194.0
416	187.63	187.63	187.09	187.99	187.4
448	179.27	179.27	178.87	179.82	178.9
480	170.00	170.00	169.64	170.21	170.0
512	160.73	160.73	160.11		164.0

Table 3. Instantaneous Frequency Estimates.

The best solution by Carmichael and Wiley (1978) uses a special zero crossing method that is applicable only to pure sine waves so that it will not generalize to more general spectra.

The adaptive CVA method did about as well as the best of the methods other than Wiley and Carmichael, and much better than a lot of them. Note that the adaptive CVA approach makes no assumptions about the form of the spectrum or the character of the time variation. Also the adaptive CVA method is completely automatic, and in this example did not involve any considerations by an analysis to determine the choice or modification of the computations. As measured in terms of the estimated instantaneous frequency, the method did very well.

5.3 Adaptation to Abrupt Changes

The primary problem in adaptive time series analysis is determining if and when an abrupt change has occurred. This problem reduces to comparing two intervals of data and determining if the same process model is a better description of the observations than a different model for each interval of data. A complication of the problem is that the exact time is unknown and must be determined from the data. This involves the comparison of a multitude of hypotheses concerning the possible time of occurrence of the abrupt change. In addition, the ability to detect

the abrupt change depends upon the data length of the data intervals used. The best data length for detection depends upon the type of abrupt change since some changes affect the observations immediately and the effect decreases rapidly, while for other changes the effect on the observations takes some time before it is apparent. Thus the consideration of different data lengths requires additional hypotheses to be considered and compared. The predictive inference and entropy methods give a sound basis for the comparison of the multitude of nonnested hypotheses.

Consider the problem of determining if there is a change in the process between two disjoint data subintervals. The detection problem considered is where the process is modeled as a slowly changing process using some efficient procedure such as given in the previous section. The notation of the previous sections will be used with subscript 1 or 2 corresponding to the data intervals Y_1 from the past slowly changing models and Y_2 from the new data that is to be compared for detection of an abrupt change. The subscripted parameters θ_1 and θ_2 with or without other superscripts, hats, or tildes will denote models based upon the corresponding data interval. The data lengths of the two intervals Y_1 and Y_2 are generally different with the first data interval determined by the slow adaptation method and with the second set usually much shorter and of variable data length since the best data length for detection of abrupt changes is not known. For any selection of the two intervals, we wish to determine if there has been a significant departure in the process characteristics between the two data sets.

Ideally, the hypothesis (θ_1, θ_2) that the models are different over the two subintervals Y_1 and Y_2 verses the hypothesis θ_j that the model is the same over the joint data set (Y_1, Y_2) would be compared. But this would involve a considerable number of comparisons. To avoid such numerous comparisons, consider the following approximation. Let data set Y_1 be chosen as the most recent optimal length interval preceding Y_2 with corresponding model θ_1 which provides a near optimal prior model Y_1 . To detect any abrupt changes in the system, consider the approximation of using the model θ_1 as an approximation to the joint model $\hat{\theta}'$.

As discussed in Section 4.3, consideration can be limited to conditional models given the past or equivalently the initial state at the beginning of the subinterval. Using such conditional models, the likelihood function reduces to

$$p(Y_1, Y_2, \theta_1) = p(Y_1, \theta_1)p(Y_2, \theta_1) \quad (5-12)$$

The model on the first data set Y_1 is the same for both the above hypothesis and the change hypothesis $p(Y_1, \theta_1)p(Y_2, \theta_2)$. The entropy measure is the difference of the above two log likelihoods which involves only the likelihoods on the second subinterval Y_2 so that

$$R(\theta_2, \theta_1) = E[\log p(Y_2, \theta_1) - \log p(Y_2, \theta_2)] \quad (5-13)$$

If the system in fact had an abrupt change between Y_1 and Y_2 , then since the data length Y_1 is much longer than Y_2 , most of the information in detecting the abrupt change is in the comparison of the two models θ_1 and θ_2 on the data Y_2 .

The problem is now to obtain an estimate of the entropy measure (4-13) from the observational data. The observed log likelihood is used as an estimate of the entropy as in (3-7). The bias of this estimate of the entropy measure is

$$\begin{aligned} E[l(Y_1) - l(Y_2)] &= E[l(\hat{\theta}) - l(Y_2)] - E[l(\hat{\theta}) - l(Y_1)] \\ &= -\dim(\theta^2) + R(\hat{\theta}, Y_2) - R(\hat{\theta}, Y_1) \end{aligned} \quad (5-14)$$

where the term $\dim(\theta^1)$ in (3-8) is not present since the estimate $\hat{\theta}^1$ is a function only of the sample Y_1 which is conditionally independent of the sample Y_2 . Thus an unbiased estimate of the difference of negentropies $R(\hat{\theta}, Y_2) - R(\hat{\theta}, Y_1)$ of the two models is

$$l(Y_1) - l(Y_2) + \dim(\theta^2) \quad (5-15)$$

This gives a test for the occurrence of an abrupt change between the two data intervals. Depending upon the nature of the change and the process characteristics, the best detection interval will vary. Some changes give most of the information about the change over a short interval while others have a cumulative effect and require a long time interval to detect.

Consider as an example of the procedure for detecting abrupt changes the ARMA(4,3) model (4-8). Three types of abrupt changes were simulated including an abrupt change in the dynamics, in the state, and in the variance of the excitation noise w_t . The results of the procedure for detecting abrupt changes for the case of no change and cases of a simulated abrupt change in the dynamics, the excitation noise, and the state are shown in Tables 4, 5, 6, and 7 respectively. In each case, the entropy measure for detecting the abrupt change was computed over various intervals of data of lengths 50, 100, and 200 samples and the abrupt change occurred at the sample time 325. In the case of no abrupt change, the entropy measure shown in Table 4 is on the average 0.5186 with a standard deviation of 0.016. As shown in Table 5, the largest value of the entropy measure is in fact in the interval samples 300-350 containing the time of the abrupt change in dynamics. The following interval of samples 350-400 also indicates a large value of the entropy measure. The initial large value in interval 300-350 is due to a transient in the state as it settles to a new steady state variance which is largely complete in interval 350-400. The entropy measure then persists at this value in following intervals. The abrupt change in noise variance shown in Table 6 has a different character. The negentropy changes abruptly in interval 300-350 and remains at that value in succeeding intervals. With an abrupt change in the state

SAMPLE TIMES	NUMBER OF POINTS IN DATA SUBINTERVAL		
	50	100	200
200			
	0.495	0.503	0.518
250	0.513		
300	0.540	0.534	
350	0.527		
400			

Table 4. Value of Per Sample AIC for Subintervals of the Sample with No Abrupt Change.

SAMPLE TIMES	NUMBER OF POINTS IN DATA SUBINTERVAL		
	50	100	200
200			
250	0.495	0.503	10.214
300	0.513		
350	28.220	19.924	
400	11.628		

Table 5. Value of Per Sample AIC for Subintervals of the Sample with an Abrupt Change in Dynamics at Sample 325.

shown in Table 7, the departure is largely confined to the interval 300-350 although the transient has not quite died out in the interval 350-400.

In all cases there was no assumption as to the nature of the change, and the procedure works as well on state jumps, changes in noise variances, or other changes. Note that the character of abrupt change is quite different depending on the type of abrupt change that occurs. The best detection of abrupt changes can only be achieved by an adaptive procedure that considers the multitude of data intervals and selects a near optimal data length for detection. These initial results on the adaptive detection procedure demonstrate that it is very sensitive to a variety of different abrupt changes in the model.

SAMPLE TIMES	NUMBER OF POINTS IN DATA SUBINTERVAL		
	50	100	200
200			
	0.495	0.503	0.579
250	0.513		
300	0.653	0.653	
350	0.654		
400			

Table 6. Value of Per Sample AIC for Subintervals of the Sample with an Abrupt Change in Excitation Noise Variance at Sample 325.

SAMPLE TIMES	NUMBER OF POINTS IN DATA SUBINTERVAL		
	50	100	200
200			
	0.495	0.503	21.842
250	0.513		
300	85.766	43.180	
350	0.595		
400			

Table 7. Value of Per Sample AIC for Subintervals of the Sample with an Abrupt Change in State at Sample 325.

6. SMALL SAMPLE MULTIVARIATE ANALYSIS

The approach to small sample inference in this Chapter is the use of the entropy measure of model approximation error to evaluate the performance of small sample methods. This general approach is based on the justification of entropy as the natural measure of model approximation error as developed in Chapter 2. The historical Bayesian predictive inference approach plays a major role in providing a computable predictive density which is subsequently shown to be optimal in terms of the entropy measure. This optimality is established by considering best invariant predictive densities. For the multivariate normal family, several predictive densities are compared with the best invariant to show the large improvements that are possible in small samples.

6.1 Bayesian Predictive Inference

The historical approach to predictive inference involves the use of Bayesian concepts and methods to determine the predictive density. Consider the parameterized class of probability density functions

$$F = \{p(y, x | \theta); \theta \in \Theta\} \quad (6-1)$$

defined on the joint sample space (X, Y) . The predictive inference setup as in Section 2.1 is considered where a predictive density $p_a(y | x)$ is to be constructed as an approximation to the true conditional density $p(y | x, \theta)$ for the unknown parameter value θ . In the Bayesian approach, the predictive density is constructed on the basis of an assumed prior density $p(\theta)$ on the parameters θ using Bayes rule.

From Bayes Theorem, the posterior density of the parameters is given by

$$p(\theta | x) = \frac{p(\theta)p(x | \theta)}{p(x)} \quad (6-2)$$

where the marginal density of x is

$$p(x) = \int p(\theta)p(x | \theta)d\theta \quad (6-3)$$

The Bayesian predictive density $p_b(y | x)$ is then given by computing the marginal density using the posterior so

$$p_b(y | x) = \int p(y | x, \theta)p(\theta | x)d\theta \quad (6-4)$$

This approach is direct and simple although the assumption of a prior density $p(\theta)$ on the

parameters θ is bothersome from both a theoretical and practical point of view.

A major objection to the Bayesian approach is the use of an arbitrary prior density on the parameters to express ignorance. If a uniform density is used for the prior on θ , then a nonsingular transformation of the parameters to a new set $\phi = g(\theta)$ and use of a uniform prior on ϕ in general produces a different posterior density. Thus there is a certain arbitrary choice of the parameterization and resulting posterior. A way around this is the use of *noninformative* prior densities. Such densities give posterior densities that are invariant to transformations of the parameter space (Jeffreys, 1961; Box and Tiao, 1973). In situations where a noninformative prior exists, it can be obtained in terms of the Fisher Information matrix.

In recent years, some intriguing connections between the Bayesian predictive density and concepts of entropy, frequentist methods, invariant methods, and noninformative priors have been made. Still in a strictly Bayesian context, consider the negentropy measure (2-6) applied to the Bayesian predictive density (6-4). Of course a Bayesian would take expectation of this measure with respect to the unknown parameters θ which will be called the *Bayes Risk*. Using (6-4) and interchanging the order of integration, the Bayes Risk between any two predictive densities $p_1(y|x)$ and $p_2(y|x)$ is

$$\begin{aligned} E_{\theta} E_{x| \theta} \int_{y|x, \theta} (p_1(y|x) ; p_2(y|x)) &= \int p(\theta) \int p(x|\theta) \int p(y|x, \theta) \log \frac{p_1(y|x)}{p_2(y|x)} dy \\ &= \int p(x) dx \int p_{\theta}(y|x) \log \frac{p_1(y|x)}{p_2(y|x)} dy \end{aligned} \quad (6-5)$$

Now setting $p_1(y|x) = p_{\theta}(y|x)$ guarantees that the Bayes Risk between p_{θ} and any predictive density p_2 is nonnegative and zero if and only if $p_2(y|x) = p_{\theta}(y|x)$.

Thus in a Bayesian context, the Bayesian predictive density is optimal in terms of the Bayes Risk, i.e. the expected negative entropy measure with the expectation also taken over the parameters θ . As was noted by Aitchison in the original derivation of this result for the case of x and y independent, there are a number of interesting cases where the negative entropy measure, i.e. the Bayes Risk excluding expectation over the parameters θ , is not a function of the parameters θ . In such cases the Bayesian predictive density is optimal in a frequency sampling sense where there is a fixed unknown true parameter value and the negative entropy (2-6) is used as the measure of error. This topic is discussed in the next section.

6.2 Best Invariant Predictive Densities

In several particular cases, the optimal predictive density has been found that minimizes the negative entropy. Murray (1977) considers the class of d -dimensional multivariate normal densities $N_d(\mu, \Sigma)$ with

$$p(y | \mu, \Sigma) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(y-\mu)^T \Sigma^{-1}(y-\mu)\right\} \quad (6-6)$$

In this case Aitchison and Dunsmore (1975, p. 29) show that using the noninformative prior density proportional to $|\Sigma|^{-1}$, the Bayesian predictive density is the d -dimensional Student distribution

$$p_p(y | x) = St_d[n-1, \hat{\mu}, (n+1)(n-1)^{-1}\hat{\Sigma}] \quad (6-7)$$

where the d -dimensional vector z is $St_d(k, b, c)$ if it has density function

$$p(z) = \frac{\Gamma\{(k+1)/2\}}{\pi^{d/2} \Gamma\{(k-d+1)/2\} |kc|^{-1/2} \{1+(z-b)^T (kc)^{-1}(z-b)\}^{(k+1)/2}} \quad (6-8)$$

This Bayes predictive density was shown (Aitchison, 1975) to result in the negative entropy not a function of the unknown parameter θ_0 . It is thus also optimal in the frequency sense.

This same result was derived by Murray (1977) using invariance concepts. Consider the class G of invariant predictive densities $p_a(y | x)$ that are invariant to translations and linear transformations of the sample x . In this class, the best invariant predictive density $p_I(y | x)$ was shown to be (6-7) which gives a constant value of the negative entropy independent of the value of the true parameter value θ_0 . This gives a strictly frequentist interpretation of the Bayes predictive density. A stronger result reported very recently (Levy and Perng, 1986) is the minimality the negentropy for the best invariant predictive density $p_I(y | x)$ uniformly in the unknown parameters $\theta_0 = (\mu_0, \Sigma_0)$ among any predictive density in the class G of invariant predictive densities.

6.3 Comparison of Entropy for Multivariate Normal

To illustrate the usefulness of the predictive inference approach using negentropy, the results for the multivariate normal distribution are given below in terms of the relative odds of the likelihood ratio. Consider the case (6-6) of the multivariate normal density $N_d(\mu, \Sigma)$. Here three methods are compared: the *estimator method* using the predictive density $p_E(y | x) = N_d(y, \hat{\mu}(x), \hat{\Sigma}(x))$ where the maximum likelihood estimates $\hat{\mu}(x)$ and $\hat{\Sigma}(x)$ are used, the *best normal* with estimates $(\hat{\mu}, (n+1)(n-d-2)^{-1}\hat{\Sigma})$ which minimize the negentropy in the class of normal densities, and the Bayesian predictive density which is identical to the best invariant

predictive density.

The expected negentropies are shown in Table 8 for the above three predictive densities.

Predictive Densities	Number of Observations (n)					
	4	6	11	14	20	50
1-Dimensional						
Estimative	1.191 (2.48)	0.366 (1.170)	0.130 (1.037)	0.094 (1.121)	0.060 (1.009)	0.021 (1.001)
Best Normal	0.476 (1.21)	0.226 (1.048)	0.103 (1.009)	0.079 (1.006)	0.053 (1.002)	0.020 (1.000)
Best Invariant	0.282	0.180	0.094	0.083	0.051	0.020
8-Dimensional						
Estimative	-	-	36.87 (10^{14})	8.08 (221.4)	2.85 (3.819)	0.61 (1.127)
Best Normal	-	-	6.79 (8.93)	3.15 (1.56)	1.63 (1.127)	0.50 (1.010)
Best Invariant	-	-	4.60	2.69	1.51	0.49

Table 8. Expected Negative Entropy (and the Geometric Mean of the Likelihood Odds Relative to the Best Invariant).

In comparing two predictive distributions, the relevant quantity is the difference between their negentropies (2-5) (Larimore (1983a)). The exponential of this quantity is the geometric mean of the relative odds of a sample y having come from the two respective predictive distributions. This exponential of the negentropy difference is also given in parentheses in Table 8 for the best

normal and estimative methods relative to the best invariant method. Since $\exp(a-b) \doteq 1+(a-b)$ for $a-b \ll 1$, we see that for negentropy differences much less than unity, the odds of an observed d -dimensional sample y coming from either of two predictive distributions is about equal. For the negentropy difference near unity, these odds are disproportionate of order $e=2.7$; and if it is much greater than unit the odds can get very large. Note that a 20 percent increase in the negentropy as between the estimative and best invariant for $d=1$ and $n=20$ has only a one percent odds advantage. On the other hand, a 17 percent increase in the negentropy as between the best normal and best invariant for $d=8$ and $n=14$ has an odds ratio of 1.56. This emphasizes the importance of comparing the negentropy on the basis of the arithmetic difference and not the relative proportion. Note that for very small samples the relative odds ratio can be much larger than unity and even in the tens or hundreds. Thus there is a huge potential gain in the use of predictive inference in very small samples as has been noted in different terms by Aitchison and Dunsmore (1975, p. 231), Aitchison and Kay (1975) and Murray (1979).

7. CONCLUSIONS AND RECOMMENDATIONS

7.1 Conclusions From Phase I Study

In this Phase I SBIR study, statistical methods are developed using predictive inference and entropy. This approach has a strong intuitive appeal as a result of the justification of the entropy measure based upon the predictive inference framework and the fundamental statistical principles of sufficiency and repeated sampling. This approach applies to a wide class of inference problems including:

- general inference methods such as parametric or nonparametric methods
- exact evaluation of small sample procedures
- determination of model order or structure including the case of non-nested multiple comparison
- time series analysis including definition of optimal tracking of time varying processes and optimal detection of abrupt changes.

The entropy measure provides a fundamental measure for the comparison of alternative statistical procedures and provides a basis for developing optimal statistical inference methods.

In this study a number of particular topics were addressed from the predictive inference and entropy perspective including:

- statistical model building involving the determination of parametric model structure and order in the general case of constrained multiple nonnested alternatives,
- time series modeling and forecasting involving the determination of parametric model structure and order,
- adaptive time series analysis involving optimal methods for tracking slow changes as well as for detecting abrupt changes or failures,
- small sample inference for multivariate distributions of the exponential family.

Some major results were developed on these topics that demonstrate the feasibility and desirability of developing statistical methods using predictive inference and entropy.

A number of results were obtained for the nonnested multiple comparison problem based upon the study of constrained maximum likelihood estimates. These include:

- consideration of the general constrained case
- general extension of Akaike's AIC procedure to constrained non-nested multiple comparison problems
- solution of the general constrained case requires that a condition on the Fisher information and Hessian matrices be satisfied
- a general model order and structure selection method was shown to be asymptotically optimal.

Previous developments consider only the case where the true parameter is approached asymptotically and exclude the case where the true parameter lies outside the models considered. The constrained case investigated in this study gives a basis for viewing the predictive inference and entropy method as model approximation when the models are restricted and asymptotically biased. These results provide a basis for the use of predictive inference and entropy on the general time series analysis and adaptive time series analysis problems involving constrained non-nested multiple comparison.

Using currently available methods, the time series analysis problem is difficult because the parametric model structure is unknown and requires the fitting and comparison of many different models. Also current methods are numerically and statistically illconditioned for some models. The approach of predictive inference and entropy provides a natural solution to the multiple comparison problem. The results obtained using the predictive inference and entropy approach for multivariate time series analysis include:

- Explicit expressions for a lower bound on the achievable accuracy in the estimation of the transfer function and power spectral matrix
- This lower bound applies to the case where the true model order is unknown and a model order determination procedure is used.
- The lower bound is achieved for large samples using maximum likelihood estimation and the AIC order determination procedure.

An example of the estimation accuracy of a true ARMA(4,3) process using spectral smoothing, AR model, and ARMA model fitting show the considerable difference in using these various methods.

Markov models of time series were developed as a basis for stable time series analysis methods using the canonical variate method (CVA). This method is numerically and statistically stable and has been applied recently to a number of high order multivariable time series analysis problems. This approach provides the basis for adaptive time series analysis methods.

Markov models of time series with changing characteristics were developed including slowly and abruptly changing processes. Using the CVA method as the computational method, the entropy methods were applied to adaptive time series analysis:

- Statistical methods for determining the optimal data length for adaptation to slow changes were developed.
- Statistical methods for choosing the optimal time interval for detection of an abrupt change in the process were developed.
- The entropy measure is optimally sensitive to any abrupt changes including changes in process dynamics, changes in the excitation noise levels, and jumps in the process state.

These results demonstrate the feasibility of developing adaptive time series analysis methods based upon entropy methods and the CVA computations. The CVA computations have been demonstrated in real time identification of multivariable systems. Thus the feasibility of adaptive time series analysis in real time has been demonstrated.

Small sample methods were developed using the predictive inference and entropy methods. The justification of entropy based upon the sufficiency and repeated sampling principles provides a sound justification for the use of recently developed small sample methods based upon entropy. The Bayesian method was extended to the case where the informative and predictive experiments are dependent. The theory is illustrated for the multivariate normal distribution using the Bayesian, best invariant, estimative, and best normal predictive densities. The relative measure of approximation is shown to be the per sample relative odds ratio which is the exponential of the entropy measure.

7.2 Recommendations for Further Research and Development

This study has demonstrated the feasibility and usefulness of predictive inference and entropy methods particularly in the areas of:

- constrained nonnested multiple comparison of models
- model order and structure determination for time series
- modeling of changing processes using Markov model structures
- optimal adaptation to slowly varying processes by optimal selection of data interval
- optimal adaptation to abrupt changes of unknown type at unknown times by optimal selection of the detection data interval
- automatic stable computation of time series models using the CVA method
- determination of lower bounds for the estimation of transfer functions and power spectra.

These achievements provide a bases for the further research and development of predictive inference and entropy methods.

The areas of greatest promise appear to be those of adaptive and nonadaptive time series analysis for the following reasons:

- the number of potential applications to DoD systems is very large
- time series analysis and adaptation are the major problems in adaptive control
- to address the adaptive time series analysis problem requires an approach that deals with the multiple comparison problem in a fundamental way that is offered by predictive inference and entropy methods
- among the current time series analysis methods, only the CVA method is suitable for real time solution of the problem
- present and near future computers are capable of multivariable identification in less than a second of computation for high order systems of dozens of states

These methods have been demonstrated to be feasible, and the development of online adaptive time series analysis software for general application would provide an enormous capability for DoD systems. Presently there are no other known approaches that will achieve this goal. Such adaptive time series processors would allow for the adaptation of systems to slow and abrupt changes in the environment.

The topics recommended for further research and development include:

- further research and development on the adaptive time series analysis methods for adaptation to slow and abrupt changes
- development of algorithms for implementation of the adaptive methods that are numerically stable and accurate and are statistically reliable
- prototype algorithm testing to demonstrate the accuracy, reliability, and computational requirements on typical DoD problems.
- software development to provide modular, documented, and verified software in one or more general programming languages.

The achievement of these objectives would provide a dramatic improvement in the performance of adaptive methods and the availability of software for adaptation in DoD systems.

REFERENCES

- Adamjan, V.M. , D.Z. Arov and M.G. Krein, (1978). Infinite Hankel Block Matrices and Related Extension Problems. *Amer. Math. Soc. Transl., series 2*, Vol. 111, 133-156.
- Aitchison, J. (1975). "Goodness of Prediction Fit," *Biometrika*, Vol. 62, pp. 547-54.
- Aitchison, J. and I.R. Dunsmore (1975). *Statistical Prediction Analysis*, Cambridge University Press.
- Akaike, H. (1976). "Canonical Correlation Analysis of Time Series and the Use of an Information Criterion." *System Identification: Advances and Case Studies*, R. K. Mehra and D.G. Lainiotis, Eds., New York: Academic Press, pp. 27-96.
- Akaike, H. (1975). "Markovian Representation of Stochastic Processes by Canonical Variables." *SIAM J. Contr.*, Vol. 13, pp. 162-173.
- Akaike, H. (1974a). "Stochastic Theory of Minimal Realization." *IEEE Trans. Automat. Contr.*, Vol 19, pp. 667-674.
- Akaike, H. (1974b). "A New Look at Statistical Model Identification." *IEEE Automatic Control*, Vol 19, pp. 667-674.
- Akaike, H., (1973). "Information theory and an extension of the maximum likelihood principle." In *2nd International Symposium on Information Theory.*, Eds. B.N. Petrov and F. Csaki, pp. 267-281. Budapest: Akademiai Kiado.
- Anderson, T.W. (1958). *An Introduction to Multivariable Statistical Analysis*. New York: John Wiley.
- Astrom, K.J. (1973). "On self-tuning regulators." *Automatica*, Vol 9, pp. 185.

- Astrom, K.J. (1983). "Theory and Applications of Adaptive Control-A Survey." *Automatica*, Vol 19, pp. 471-86.
- Astrom, K.J., U. Borisson, L. Ljung and B. Wittenmark, (1977). "Theory and applications of self-tuning regulators." *Automatica*, Vol 13, pp. 457.
- Astrom, K.J. and B. Wittenmark (1973). "On self-tuning regulators", *Automatica*, Vol 9, pp. 185.
- Bhansali, R.J., and Downham, D.Y. (1977), "Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion." *Biometrika*, Vol 64, pp. 547-71.
- Box, G.E.P. and G.M. Jenkins (1970), *Time Series Analysis Forecasting and Control*, San Francisco: Holden-Day
- Box, G.E.P., and G.C. Tiao (1973). *Bayesian Inference in Statistical Analysis*. Reading, MA: Addison-Wesley.
- Camuto, E and Menga, G (1982), "Approximate Realization of Discrete Time Stochastic Processes," *Proceedings of the Sixth IFAC Symposium on Identification and System Parameter Estimation*, Vol 2, Eds G.A Bekey and G.N. Saridis, Washington, D.C., June, 1982.
- Candy, J.V., Bullock, T.E., and Warren, M.E. (1979), "Invariant Description of the Stochastic Realization," *Automatica*, Vol. 15, pp. 493-5.
- Carmichael, W.R., and R.G. Wiley (1978). "Instantaneous Frequency Estimation from Sampled Data," *Proceedings of the RADC Spectrum Estimation Workshop*, held May 24-6, 1978, at Rome Air Force Base, Rome, NY. Technical Report No. AD-A054650, Defense Technical Information Center, Alexandria, VA.
- Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman and Hall.
- Desai, U.B., and D. Pal (1982), "A Realization Approach to Stochastic Model Reduction and Balanced Stochastic Realization," *Proc. 21st Conf. on Decision and Control*, Orlando, pp. 1105-12.

- Evans, R.J. and R.E. Betz (1982). "New Results and Applications of adaptive control to classes of nonlinear systems." *Proc. Workshop on Adaptive Control*. Florence, Italy.
- Friedlander, B. (1982). "Lattice Filters for Adaptive Processing," *Proc. IEEE*, Vol 70, pp. 829-67.
- Friedlander, B. (1982). "Lattice Methods for Spectral Estimation," *Proc. IEEE*, Vol 70, pp. 990-1017.
- Friedlander, B. (1983). "Lattice Implementation of Some Recursive Parameter- Estimation Algorithms," *Int. J. Control*, Vol. 37, pp. 661-684.
- Gerhardt, L.A. (1978), "A Comparison of Spectrum Estimation Techniques Using Common Data Sets," *Proceedings of the RADC Spectrum Estimation Workshop*, held May 24-6, 1978, at Rome Air Force Base, Rome, NY. Technical Report No. AD-A054650, Defense Technical Information Center, Alexandria, VA.
- Gersch, W., and D.R. Sharp (1973). "Estimation of Power Spectra with Finite-Order Autoregressive Models," *IEEE Trans. Automatic Control*, Vol. 18, pp. 367-79.
- Gevers, M. and V. Wertz (1982), "On the Problem of Structure Selection for the Identification of Stationary Stochastic Processes," *Proc. of the IFAC Symp. on Identification and System Parameter Estimation*, G. Bekey and G. Saridis (eds.), Wash. D.C.: McGregor-Werner, pp. 387-92.
- Golub, G.H. (1969). *Matrix Decompositions and Statistical Calculations*. *Statistical Computation*, R.C. Milton and J.A. Nelder, eds., New York: Academic Press, pp. 365-379.
- Hagglund, T. (1983). "New Estimation Techniques for Adaptive Control." Report CODEN:LUTFD2/(TFRT-1025)1-120/(1983), Department of Automatic Control, Lund Institute of Technology. Doctoral Dissertation.
- Hart, P.E. (1971), "Entropy and Other Measures of Concentration," *J. Roy Statist. Soc., A*, Vol 134, pp. 73-85.

- Honig, M.L., and D.G. Messerschmitt (1984). *Adaptive Filters: Structures, Algorithms, and Applications*. Boston: Kluwer Academic Publishers.
- Hotelling, H. (1936). "Relations between Two Sets of Variates." *Biometrika*, Vol 28, pp. 321-377.
- Irving, E. (1979). "New Developments in improving power network stability with adaptive control." *Proc. Workshop on Applications of Adaptive Control*. Yale University, New Haven.
- Isermann, R. (1984). "Process Fault Detection Based on Modeling and Estimation Methods - A Survey," *Automatica*, Vol. 20, pp. 387-404.
- Jeffreys, H. (1961). *Theory of Probability*, Clarendon Press.
- Kendall, M.G. (1973). "Entropy, Probability and Information," *International Statistical Review*, Vol 41, pp. 59-68.
- Kullback, S. (1959), *Information Theory and Statistics*, Dover.
- Kullback, S. and R.A. Leibler (1951), "On Information and Sufficiency," *Ann. Math. Statistics*, 22, pp. 79-86.
- Kung, S.Y. and D.W. Llin (1981). "Optimal Hankel-Norm Model Reductions: Multivariable Systems", *Trans. Aut. Control*, Vol 26, pp. 832-852.
- Larimore, W.E. and R.K. Mehra (1984), "Technical Assessment of Adaptive Flutter Suppression Research," Air Force Wright Aeronautical Lab Report No. AFWAL-TR-84-3052, SSI.
- Larimore, W., S. Mahmood and R.K. Mehra (1983). "Adaptive Model Algorithmic Control." *Proc. IFAC Workshop on Adaptive Systems in Control And Signal Processing*, San Francisco, CA, June 1983.
- Larimore, W.E. (1986a). "Optimal Adaptive Identification of Changing Systems," Draft.

- Larimore, W.E. (1986b). "A Unified View of Reduced Rank Multivariate Prediction Using a Generalized Singular Value Decomposition," Submitted for publication in *Annals of Statistics*.
- Larimore, W.E. (1986c). "Simultaneous Confidence Bands for Efficient Parametric Spectral Estimation," Submitted for publication in *Biometrika*.
- Larimore, W.E. (1986d). "Constrained Nonnested Multiple Comparison Using Predictive Inference and Entropy," Draft.
- Larimore, W.E. (1986e). "Achievable Accuracy in Parametric Estimation of Multivariate Spectra". Draft.
- Larimore, W.E. (1983a). "Predictive inference, sufficiency, entropy, and an asymptotic likelihood principal." *Biometrika*, Vol 70, pp. 175-81.
- Larimore, W.E. (1983b). "System Identification, Reduced-Order Filtering and Modeling Via Canonical Variate Analysis." *Proc. 1983 American Control Conference*, H.S. Rao and T. Dorato, eds., New York: IEEE. pp. 445-51.
- Larimore, W.E. (1982). "A survey of some recent developments in system parameter identification." *Proceedings of the 6th IFAC Symposium on Identification and System Parameter Estimation*, Vol 1. Washington, D.C., June 1982, pp. 979-84.
- Larimore, W.E. (1981a). "Small sample methods for maximum likelihood identification of dynamical processes." *Applied Time Series Analysis, Proceedings of the Fifth International Time Series Meeting*. Houston, Texas, August 1981. Amsterdam, North Holland, pp. 167-174.
- Larimore, W.E. (1981b). "Recursive maximum likelihood and related algorithms for parameter identification of dynamical processes." *Proceedings of the 20th IEEE Conference on Decision and Control*, Vol 1, pp. 50-55, San Diego, California, December 1981.
- Larimore, W.E. (1977). "Nontested Tests on Model Structure." *Proceedings Joint Automatic Control Conf.* (San Francisco, CA). New York: IEEE, pp. 686-690.

- Larimore, W.E. (1977b). "Statistical inference on stationary random fields." *Proc. IEEE*, Vol 65 , pp. 961-70.
- Levy, S., and S.K. Perng (1986). "An Optimal Prediction Function for the Normal Linear Model," *J. Amer. Statist. Assoc.*, Vol. 81, pp.196-8.
- Lindquist, A. and M. Pavon (1981). "Markovian Representation of Discrete-Time Stationary Stochastic Vector Processes," *Trans. IEEE Conf. Decision and Control*, vol. 3, San Diego, pp. 1345-1356.
- Lou, X.C., A.S. Willsky, and G.C. Verghese (1983). "Failure with Uncertain Models," *Proc. Amer. Control Conf.* San Francisco, California.
- Ljung, L. and T. Soderstrom (1983). *Theory and Practice of Recursive Identification*. Cambridge: MIT Press.
- Ljung, L., I. Gustavsson and T. Soderstrom (1974). "Identification of Linear Multivariable Systems Operating Under Linear Feedback Control." *IEEE Trans. Auto. Control*, AC-19, pp. 836-840.
- Mehra, R.K. (1978). "A Survey of Time-Series Modeling and Forecasting Methodology." *Modeling, Identification and Control in Environmental Systems*, E. Vansteenkiste, ed., North Holland Publishing Co.
- Mehra, R.K. and A. Cameron (1980). "Handbook on Business and Economic Forecasting for Single and Multiple Time Series." Scientific Systems, Inc., Notes for the Institute of Professional Education Seminar.
- Mehra, R.K. and A. Cameron (1976). "A Multidimensional Identification and Forecasting Technique Using State Space Models." ORSA/TIMS Conf. Miami, FL, November 1976.
- Mehra, R.K. and J.S. Tyler (1973). "Case Studies in Aircraft Parameter Identification." *Proc. 3rd IFAC Conf. on Identification and System Parameter Estimation*, P. Eykhoff, ed., Oxford: Pergamon Press, 117-144.

- Mehra, R.K. and J. Peschon (1971). "An innovations approach to fault detection and diagnosis in dynamic systems." *Automatica*, Vol 7, pp. 657.
- Murray, G.D. (1977), "A note on the estimation of probability density functions." *Biometrika*, Vol 64, pp. 150-2.
- Murray, G.D. (1979), "The estimation of multivariate normal density functions using incomplete data." *Biometrika*, Vol 66, pp. 375-80.
- Schwartz, G. (1978), "Estimating the Dimension of a Model," *Ann. Statistics*, Vol 6, 2, pp. 461-4.
- Shibata, R., (1983). "A Theoretical View of the Use of AIC," *Proc. of the International Time Series Meeting*, O.D. Anderson, Ed.
- Shibata, R., (1981a). "An optimal autoregressive spectral estimate." *Ann. Statist.* Vol 9, pp. 300-6.
- Shibata, R., (1981b). "An optimal selection of regression variables." *Biometrika*, Vol 68, pp. 45-54.
- Wellstead, P.E. and S.P. Sanoff, (1981). "Extended self-tuning algorithm." *Intl. J. Control*. Vol 34, pp. 433-455.
- Willsky, A.S., (1980). "Failure Detection in dynamic systems." AGARD, No. 109.
- Willsky, A.S., (1976). "A Survey of Design Methods for Failure Detection Systems," *Automatica*, Vol 12, pp. 601-611.
- Willsky, A.S. and H.L. Jones (1974). "A Generalized likelihood ratio approach to state estimation in linear systems subject to abrupt changes." *Proc. IEEE Conf. on Decision and Control*. Phoenix, Arizona.

APPENDIX A

CONSTRAINED NONNESTED MULTIPLE COMPARISON USING PREDICTIVE INFERENCE AND ENTROPY

By Wallace E. Larimore

Scientific Systems Inc., Cambridge, Massachusetts, U.S.A.

SUMMARY

Research Sponsored by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Contract Number F49620-85-C-0086.

Paper in preparation for submission.

1. Introduction

The general problem of choosing a model from among a multitude of alternative models remains one of the difficult problem of statistical inference. Traditional methods of statistical hypothesis testing apply directly only to the case where there are two hypotheses under consideration and one is a subset of the other, i.e., the two hypotheses are nested. In such cases, classical methods are applicable and lead to well understood results. In the case where there are more than two hypotheses involved, the use of the classical methods are not well defined or understood even in the nested case (see for example the discussion in Anderson, 1971, pp. 270). Although the probability of rejecting one hypothesis in comparing any pair is well defined, the repeated application of such pairwise comparisons results in a test whose properties are not understood. In the case of comparing two hypotheses that are not nested, the distribution theory is available but much more complicated (Larimore, 1977).

Beyond these difficulties in carrying out the classical procedures is the lack of a general framework for formulating and solving the problem of nonnested multiple comparison of constrained models. The predictive inference approach offers a predictive measure of the accuracy of various model selection procedures that apply as easily to the case of non-nested multiple comparison. The adequacy of a model selection procedure is measured in terms of the accuracy of the selected models in hypothetical repeated "future" experiments. This is very attractive in the context of scientific inference where the role of model building is to provide a basis for prediction of the future behavior of a phenomenon. The entropy measure provides a most sensitive measure based upon the sufficient statistic as contained in the likelihood ratio. The derivation of the entropy as a measure of the prediction error of a predictive density in the predictive inference framework is based upon the fundamental principles of sufficiency and repeated sampling (Larimore, 1983). This provides a strong theoretical basis for the use of entropy in predictive settings of scientific inference. In more narrowly defined problems of quality control or decision theory involving a well defined loss function, other procedures may be more appropriate. But in a predictive scientific setting, the approach using predictive inference and entropy seems much more justified.

Consider the problem of choosing among a multitude of model structures on the basis of a set of observations. If we adopt the predictive criterion that the chosen model should be the best in a predictive sense in predicting another independent sample from

the same process, then the optimal choice is the model selection procedure with the minimum negentropy. The major problem is the practical evaluation of the negentropy measure and the determination of efficient procedures that come close to minimizing the negentropy measure.

In this paper, the theory of inference for nonnested multiple comparison is developed in the context of predictive inference and entropy. To develop a general theory, the case of maximum likelihood is considered for moderate and large samples. The case where the true process model is not contained in the models considered for inference is the usual situation in scientific inference since even the most general model forms usually do not include certain complexities such as nonlinearities, nonstationarity, etc., that may have a small effect or be very difficult to handle. Previous approaches involving the entropy measure have not explicitly included this miss-modeling. It is shown that this miss-modeling can be directly considered in the analysis. The resulting theory is very attractive in that it gives an explicit interpretation of the predictive inference approach as model approximation of the true process using simplified alternative model forms. The entropy methods lead to procedures that select models that in the predictive sense are the most accurate in approximating the true process model. The classical difficulties of nonnested and multiple comparison do not arise in this predictive inference setting.

In the paper, first the subject of constrained maximum likelihood estimation is developed in the predictive inference and entropy context. This is used to derive the expected negentropy for maximum likelihood estimates, and then to determine an unbiased estimate of the entropy. Finally, bounds on the achievable accuracy of model selection procedures is derived that depend on the number of estimated parameters in the model fitting.

2. Constrained Maximum Likelihood Estimation

In this section, properties of the maximum likelihood parameter estimates are developed for the case that the true probability model is not contained in the class of parameterized densities that are considered for inference. The classical development of the asymptotic consistency and minimum variance of maximum likelihood estimators is for the case where the true density is contained in the parametric class.

The predictive inference framework as in Larimore (1983) is adopted here with $p(x, \theta)$ the parameterized probability density where θ is a vector of parameters, x is the informative sample and y is the predictive sample. Suppose that the parameter vector $\theta^T = (\theta_1, \theta_2, \dots)$ is a finite or infinite set of parameters, and for each subset of distinct positive integers $k = (k_1, \dots, k_m)$ consider the subspace Θ_k of θ such that only the corresponding $\theta_{k_1}, \dots, \theta_{k_m}$ are nonzero where θ^k denotes a member of Θ_k , and let C_k be the class of models $C_k = \{p(x, \theta^k), \theta^k \in \Theta_k\}$. These classes of models are in general nonnested so that we do not in general have $C_k \subset C_j$ or $C_j \subset C_k$. The maximum likelihood estimator for the class C_k will be denoted as $\hat{\theta}^k(x)$.

The development of the maximum likelihood theory is straight forward for the case where Taylor series expansions are possible. This holds under the following regularity conditions (Cox and Hinkley, p. 281):

- (i) The parameter space is closed and compact.
- (ii) The probability distributions defined by any two different values of θ are distinct.
- (iii) The first three derivatives of the log likelihood $l(x, \theta)$ with respect to θ exists in the neighborhood of the true parameter value almost surely. Further, in such a neighborhood, n^{-1} times the absolute value of the third derivative is bounded above by a function of x , whose expectation exists.

In particular, these conditions permit the interchange of expectation and differentiation up to second order.

In the discussion various order models are considered, and the relationships between the various orders is developed. The log likelihood function of the informative sample x will be denoted by $l(x, \theta)$, and the gradient row vector and Hessian matrix denoted $l'(x, \theta)$ and $l''(x, \theta)$ respectively. Expectation, denoted E , will be with respect to the true density $p(x, \bar{\theta})$ unless stated otherwise where $\bar{\theta}$ denotes the true parameter value. Define the *projection* $\bar{\theta}^k$ of $\bar{\theta}$ onto Θ_k as the parameters $\theta^k \in \Theta_k$ minimizing the negentropy R_x relative to the informative sample x

$$R_x(\bar{\theta}, \theta^k) = El(x, \bar{\theta}) - El(x, \theta^k) \quad (2-1)$$

At the minimum $\bar{\theta}^k$, the gradient of (2-1) is zero so from the regularity conditions

$$El'(x, \bar{\theta}^k) = 0, \quad (2-2)$$

and the minimum is unique if and only if the expected Hessian, denoted D_x^k , of (2-1) is positive definite in an open neighborhood of the minimum. From the regularity conditions, the Hessian is given by $D_x^k = E l''(x, \bar{\theta}^k)$.

To determine the moments of the maximum likelihood estimates $\hat{\theta}^k$, consider the first order equality

$$0 = l'(x, \hat{\theta}^k) = l'(x, \bar{\theta}^k) + (\hat{\theta}^k - \bar{\theta}^k)^T l''(x, \bar{\theta}^k) \quad (2-3)$$

Taking expectation with respect to the the true density and using (2-2) gives the equation

$$D_x^k (E \hat{\theta}^k - \bar{\theta}^k) = 0 \quad (2-4)$$

that holds asymptotically for large informative sample N . For θ^k identifiable, i.e. $\bar{\theta}^k$ unique, D_x^k is nonsingular which implies that to first order

$$E \hat{\theta}^k = \bar{\theta}^k \quad (2-5)$$

Now using (2-3), the covariance of the estimation error is

$$E (\hat{\theta}^k - \bar{\theta}^k)(\hat{\theta}^k - \bar{\theta}^k)^T = (D_x^k)^{-1} E \{ l'^T(x, \hat{\theta}^k) l'(x, \bar{\theta}^k) \} (D_x^k)^{-1} \quad (2-6)$$

Note that in the unconstrained case, the middle term which is the Fisher information matrix is equal to minus the expected Hessian D_x^k , but this is not in general true for the constrained case.

3. Expected Negative Entropy for Maximum Likelihood Estimates

To evaluate the expected negative entropy for maximum likelihood estimates, consider the predictive inference setup as in Larimore (1983). The general case of dependence between the informative and predictive samples is considered. The expected negative entropy is a measure of the degree of approximation of the true conditional density $p(y|x, \bar{\theta})$ by the predictive density $p(y|x, \hat{\theta}^k)$ for predicting the future predictive sample y from the informative sample x . The expectation will be taken in two steps, first with respect to the random variable $y|x$ and then with respect to x . In this section the likelihood function $l(\hat{\theta}^k) = l(y|x, \hat{\theta}^k(x))$ is for the predictive sample $y|x$, and the maximum likelihood estimator $\hat{\theta}^k(x)$ is on the informative sample x .

Expanding (2-1) in a Taylor series gives a second order expression for the information distance which holds asymptotically for large sample size of the informative sample,

i.e. for the maximum likelihood estimate $\hat{\theta}^k$ close to the projection $\tilde{\theta}^k$

$$\begin{aligned} R_{\mathcal{M}_x}(\tilde{\theta}, \hat{\theta}^k(x)) &= E[l(\tilde{\theta}^k) - l(\hat{\theta}^k)] + E[l(\tilde{\theta}) - l(\tilde{\theta}^k)] \\ &= -E[l'(\tilde{\theta}^k)(\hat{\theta}^k - \tilde{\theta}^k)] - E\left[\frac{1}{2}(\hat{\theta}^k - \tilde{\theta}^k)^T l''(\tilde{\theta}^k)(\hat{\theta}^k - \tilde{\theta}^k)\right] + E[l(\tilde{\theta}) - l(\tilde{\theta}^k)] \\ &= -\frac{1}{2}E_x \|\hat{\theta}^k(x) - \tilde{\theta}^k\|_{D^k}^2 + R_{\mathcal{M}_x}(\tilde{\theta}, \tilde{\theta}^k) \end{aligned} \quad (3-1)$$

since $\hat{\theta}^k(x)$ is independent of $l(y|x, \tilde{\theta}^k)$ and using the gradient property of the projection (2-2). The second order expansion is only locally in the estimation error $\hat{\theta}^k - \tilde{\theta}^k$ about the projection $\tilde{\theta}^k$ of the true parameter value $\tilde{\theta}$ on the subspace of the parameters corresponding to the model θ^k . Note that this expression gives the exact bias term $R_{\mathcal{M}_x}(\tilde{\theta}, \tilde{\theta}^k)$ in small samples and involves no approximation.

4. Unbiased Estimation of Entropy

For decision on model parametric order and structure, it is necessary to estimate the negative entropy based on the informative sample. One such procedure is due to Akaike (1973). We consider the case where the informative sample x and the predictive sample y are independent. For each selection of a parameter subset $k = (k_1, \dots, k_m)$, the Akaike information criterion for comparing the maximum likelihood estimators is

$$AIC(k) = -2\log p(x, \hat{\theta}^k(x)) + 2K(k) \quad (4-1)$$

where $K(k)$ is the number of parameters, i.e. the dimension of θ^k . The minimum AIC estimator (MAICE), denoted $\hat{\theta}_A(x)$, is $\hat{\theta}_A(x) = \hat{\theta}^{\hat{k}(x)}(x)$ where $\hat{k}(x)$ is the parameter set minimizing $AIC(k)$. The $AIC(k)$ is an unbiased estimator of the negative entropy based upon the informative sample and the assumed model structure. The predictive sample is essentially replaced by the informative sample, and the term $2K(k)$ is an adjustment for the bias due to the correlation between the informative sample x and the estimate $\hat{\theta}^k(x)$.

Following Akaike, we use the maximized log likelihood $l_x(\hat{\theta}^k) = l(x, \hat{\theta}^k(x))$ as an estimate of the relative entropy and compute the bias in the procedure. We expand the log likelihood function as in (3-1) except that below the likelihood is on the informative sample so that there is dependence between $l(x, \tilde{\theta}^k)$ and $\tilde{\theta}^k(x)$. In particular, using (2-3) gives

$$-E[l'(\hat{\theta}^k)(\hat{\theta}^k - \bar{\theta}^k)] = E[(\hat{\theta}^k - \bar{\theta}^k)^T l''(\bar{\theta}^k)(\hat{\theta}^k - \bar{\theta}^k)] \quad (4-2)$$

Consider expanding the expected log likelihood difference as in (3-1) but using (4-2) as a result of the dependence. Thus

$$\begin{aligned} E[l(\bar{\theta}) - l(\hat{\theta}^k)] &= E[l(\bar{\theta}^k) - l(\hat{\theta}^k)] + E[l(\bar{\theta}) - l(\bar{\theta}^k)] \\ &= -E[l'(\bar{\theta}^k)(\hat{\theta}^k - \bar{\theta}^k)] - E\left[\frac{1}{2}(\hat{\theta}^k - \bar{\theta}^k)^T l''(\bar{\theta}^k)(\hat{\theta}^k - \bar{\theta}^k)\right] + E[l(\bar{\theta}) - l(\bar{\theta}^k)] \\ &= E[(\hat{\theta}^k - \bar{\theta}^k)^T l''(\bar{\theta}^k)(\hat{\theta}^k - \bar{\theta}^k)] + R(\bar{\theta}, \bar{\theta}^k) \\ &= -tr(D_x^k)^{-1} E\{l' T(x, \bar{\theta}^k) l'(x, \bar{\theta}^k)\} + R(\bar{\theta}, \bar{\theta}^k) \end{aligned} \quad (4-3)$$

where the third equality follows using the expression (3-1) for the expected negentropy. In the general case, the bias term in the negentropy, $R(\bar{\theta}, \bar{\theta}^k)$, is correctly estimated except for the trace term. Asymptotically, if $\hat{\theta}^k$ approaches $\bar{\theta}$, then the matrix of the trace is the identity. This is the case considered by Akaike (1973). In the general constrained case, this may not be the case so that the bias depends upon the unknown true parameter value. What is required is a restriction such as the Hessian and Fisher information matrix being equal globally which gives

$$-tr(D_x^k)^{-1} E\{l' T(x, \bar{\theta}^k) l'(x, \bar{\theta}^k)\} = tr I_{\mathbf{E}(k)} = K(k) \quad (4-4)$$

Consider the case of fitting two models $\hat{\theta}^k$ and $\hat{\theta}^j$, and consider the expected difference of the maximized log likelihoods

$$\begin{aligned} E[l(\hat{\theta}^k) - l(\hat{\theta}^j)] &= E[l(\bar{\theta}) - l(\hat{\theta}^j)] - E[l(\bar{\theta}) - l(\hat{\theta}^k)] \\ &= +dim(\theta^k) - dim(\theta^j) + R(\bar{\theta}, \hat{\theta}^j) - R(\bar{\theta}, \hat{\theta}^k) \end{aligned} \quad (4-5)$$

Thus for relative comparisons among hypotheses based on a given sample, an unbiased estimate of twice the negentropy $E[l(\bar{\theta}) - l(\hat{\theta}^k)]$ is given by the Akaike information criterion. Note that the proof of this is much more general than that originally given by Akaike (1973) since it applies to the general case of comparisons of nonnested structures. Also, the true parameter $\bar{\theta}$ need not be contained in the structures being compared so long as the Fisher information matrix is a constant in a neighborhood including the true parameter and its projection onto the subspaces of these structures.

5. Bounds on the Accuracy of Model Selection

To obtain a correction for the bias in the sample log likelihood as an estimate of the entropy measure, the Fisher information must be constant or other restrictions are required. In this section the Fisher information is assumed to be constant in a neighborhood of the true parameter $\bar{\theta}$ containing the projection $\bar{\theta}^k$. Under these conditions a lower bound on the entropy measure is derived. From the above relationships and standard arguments of asymptotic consistency, it follows (Cox and Hinkley, 1974, p. 292) that the constrained maximum likelihood estimate $\bar{\theta}^k$ is consistent and the limit in probability is $\bar{\theta}^k$. In addition the properties of the unconstrained maximum likelihood translate to the projection $\bar{\theta}^k$ since the Fisher information matrix is constant.

Consider first the case of estimating the model $\bar{\theta}$ using the k -th order maximum likelihood estimator $\bar{\theta}^k$. Then from (3-1)

$$R(\bar{\theta}, \bar{\theta}^k) = \frac{-1}{2} \text{tr } D^k E\{(\bar{\theta}^k - \bar{\theta})(\bar{\theta}^k - \bar{\theta})^T\} + R(\bar{\theta}, \bar{\theta}^k) \geq \frac{K(k)}{2N} + R(\bar{\theta}, \bar{\theta}^k) \quad (5-1)$$

where the inequality follows from the Cramer-Rao lower bound $E_x[(\bar{\theta}^k - \bar{\theta})^T (-D^k)(\bar{\theta}^k - \bar{\theta})] \geq \text{tr } I_{K(k)}/N = K(k)/N$ where $K(k)$ is the number of parameters estimated in the model $\bar{\theta}^k$ and N is the sample size. The last term in (5-1) is the bias in using too low an order in the model fitting, and the first term is the sampling variability apart from the bias. This bound $K(k)/2N$ on the variability is achieved for an asymptotically unbiased and efficient estimator for the class C_k such as maximum likelihood. In particular, if the true model order is no greater than k , then

$$\lim_{N \rightarrow \infty} N[R(\bar{\theta}, \bar{\theta}^k) - R(\bar{\theta}, \bar{\theta}^k)] \geq \frac{k}{2} \quad (5-2)$$

The true order k of the process is usually not known and may in fact be infinite, and the bias term in (5-1) is not known so that the above discussion is not very useful in practice although it does give some insight into the accuracy issue.

Consider now the Akaike MAICE procedure using the estimator $\hat{\theta}_A$. Assume that the true model order is infinite, so that for any j' there exists a $j \geq j'$ such that $\theta_j > 0$. Define the optimal predictive order $k^*(N)$ depending upon the sample size N as the order k minimizing the negentropy (5-1). Then under suitable assumptions, the remarkable result is obtained by Shibata (1981a, 1981b, 1983) that asymptotically as $N \rightarrow \infty$

- (i) the lower bound using any order selection scheme is for each N given by evaluating (5-1) at $k = k^*(N)$, and
- (ii) this lower bound is achieved by the MAICE estimator $\hat{\theta}_A$.

Thus the lower bound on the negentropy which is achieved by MAICE is equal to the negentropy that would result from using an efficient estimator with apriori knowledge of the optimal order $k^*(N)$.

References

- Akaike, H., (1973). "Information theory and an extension of the maximum likelihood principle." In *2nd International Symposium on Information Theory*, Eds. B.N. Petrov and F. Csaki, pp. 267-281. Budapest: Akademiai Kiado.
- Anderson, T.W. (1971). *The Statistical Analysis of Time Series*. New York: Wiley and Sons.
- Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman and Hall.
- Larimore, W.E. (1983). "Predictive inference, sufficiency, entropy, and an asymptotic likelihood principal." *Biometrika*, Vol 70, pp. 175-81.
- Larimore, W.E. (1977). "Nontested Tests on Model Structure." *Proceedings Joint Automatic Control Conf.* (San Francisco, CA). New York: IEEE, pp. 686-690.
- Shibata, R., (1983). "A Theoretical View of the Use of AIC," *Proc. of the International Time Series Meeting*, O.D. Anderson, Ed.
- Shibata, R., (1981a). "An optimal autoregressive spectral estimate." *Ann. Statist.* Vol 9, pp. 300-6.
- Shibata, R., (1981b). "An optimal selection of regression variables." *Biometrika*, Vol 68, pp. 45-54.

APPENDIX B

ACHIEVABLE ACCURACY IN PARAMETRIC ESTIMATION OF MULTIVARIATE SPECTRA

By Wallace E. Larimore

Scientific Systems Inc., Cambridge, Massachusetts, U.S.A.

Research Sponsored by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Contract Number F49620-85-C-0086.

Paper in preparation for submission.

1. INTRODUCTION

The problem of determining the achievable accuracy in identifying a model for a stationary multiple time series is considered in this paper. The cases of the presence or absence of an exogenous input or additive measurement noise are included. Consider the general case where $x(t)$ is the exogenous input and $y(t)$ is the observed endogeneous output of a system which may include other unknown excitations and measurement noise. Thus consider the jointly stationary gaussian vector time series $x(t)$ and $y(t)$, $t = \dots, -1, 0, 1, \dots$, with power cross-spectral matrices $S_{xx}(\omega, \theta)$, $S_{xy}(\omega, \theta)$, $S_{yy}(\omega, \theta)$ parameterized by θ , and denote the power cross-spectral matrix of the joint vector $(x^T(t), y^T(t))^T$ as $S(\omega, \theta)$.

Statistical inference is considered on a class of linear Gaussian processes parameterized by θ . Specifying a parametric model for the conditional process $y(t)$, $t \geq s$ given $x(t)$, $t < s$ implies a causal linear model of the form

$$y(t) = q(t) + \sum_{\tau=0}^{\infty} h(t-\tau; \theta) x(t) = q(t) + r(t) \quad (1)$$

where $h(t; \theta)$ is a causal linear system giving the response $r(t)$ in $y(t)$ due to the past exogeneous input $x(t)$ and where $q(t)$ is the error in predicting $y(t)$ by $r(t)$. From linear prediction theory (Gikhman and Skorokhod, 1969), the transfer function of $h(t; \theta)$ is $H(\omega; \theta) = S_{xx}^{-1}(\omega, \theta) S_{xy}(\omega, \theta)$, and the error $q(t)$ in predicting $y(t)$ is uncorrelated with $r(t)$ with power spectrum $S_{qq}(\omega; \theta) = S_{yy}(\omega, \theta) - H(\omega, \theta) S_{xx}(\omega, \theta) H^*(\omega, \theta)$. Note that any class of parameterized models $S(\omega, \theta)$ can be equivalently specified by the parameterized models $(S_{qq}(\omega, \theta), H(\omega, \theta))$ which will prove more convenient.

2. ENTROPY AND SPECTRAL ACCURACY

Consider the following predictive inference setting (Larimore, 1983) involving an observed *informative sample* $u^T = (x^T(1), y^T(1), \dots, x^T(N), y^T(N))$ of size N used to estimate the process model, and similarly consider a conceptual *predictive sample* v of size M used to evaluate the accuracy of the estimated model. The predictive sample is assumed to be identically distributed but independent of the informative sample. Consider the problem of inference on the parametric class $\{p(v, \theta), \theta \in \Theta\}$ of models with probability densities $p(v, \theta)$ based upon the informative sample u . Consider the conceptual repeated sampling experiment where on each trial the samples u and v are each drawn independently from the process $S(\omega, \theta_0)$ with θ_0 assumed to be the true parameter value. An estimative model $\hat{p} = p(v, \hat{\theta}(u))$ is chosen for the density of v by some parameter estimation scheme $\hat{\theta}(u)$. The *negative entropy*, also known as the expected Kullback-Leibler discrimination information or expected I-divergence, is a measure of the error in approximating the true density p_0 of v by the estimate \hat{p} and is given by

$$R(p_*, \hat{p}) = E_u K(p_*, \hat{p}) = E_u \int p(v, \theta_*) \log \frac{p(v, \theta_*)}{p(v, \hat{\theta}(u))} dv \quad (2)$$

where E_u denotes expectation relative to u and K denotes the Kullback-Leibler information. The negative entropy measure follows as the natural measure in the predictive inference setting from the fundamental principles of sufficiency and repeated sampling (Larimore, 1983). This approach applies to very general modeling methods such as nonparametric, semiparametric or parametric procedures as well as methods including decisions on model structure or order such as those used for AR and ARMA modeling.

Let lower case variables denote a sample of size M of the predictive sample, e.g. $y = (y^T(1), y^T(2), \dots, y^T(M))^T$ and Σ_{yy} denote the covariance matrix of y . By expressing the density $p(y, x; \theta) = p(y-r; \theta) p(x; \theta)$ in terms of the conditional random process $q(t) = y(t) - r(t)$,

$$p(y, x, \theta) = p(y|x, \theta) p(x, \theta) = p(y - r(x, \theta), \theta) p(x, \theta) = p(q|r(x, \theta), \theta) p(x, \theta) \quad (3)$$

the log likelihood separates with the density of $x(t)$ in many problems not a function of the unknown parameters or at least a function of a separate set of parameters. A conditional viewpoint is taken in the following where only the conditional term $p(q|r(x, \theta), \theta)$ with x considered as non-random is considered. The dependence of r on x will be understood in the notation. Inclusion of the second term is tantamount to modeling the joint vector time series involving the two series $x(t)$ and $y(t)$ jointly rather than as exogeneous and endogeneous respectively. The joint case is included as a special case of $\bar{y}^T(t) = (y^T(t), x^T(t))$ a vector process with no input $\bar{x}(t)$ which will be discussed as a particular instance of the model throughout the paper. The I-divergence (2) conditional on x thus becomes

$$\begin{aligned} K(p_*, \hat{p}) &= \int p(q|x, \theta_*) \log \frac{p(q|x, \theta_*)}{p(q|x, \hat{\theta})} dq = E \log p(y-r_*, \Sigma_{qq}) - E \log p(y-\hat{r}, \hat{\Sigma}_{qq}) \\ &= E \log p(y-r_*, \Sigma_{qq}) - E \log p((y-r_*) + (r_* - \hat{r}), \hat{\Sigma}_{qq}) \\ &= E \log p(y-r_*, \Sigma_{qq}) - E \log p((y-r_*), \hat{\Sigma}_{qq}) - E (r_* - \hat{r})^T \hat{\Sigma}_{qq}^{-1} (r_* - \hat{r}) \end{aligned} \quad (4)$$

where $\hat{r} = r(x, \hat{\theta})$, $r_* = r(x, \theta_*)$, and where E denotes expectation with respect to the density $p(y|x, \theta_*)$.

For brevity set S denote the true spectrum, and let \hat{S} denote an estimate of S . We will need to assume that $S(\omega)$ is continuous and that $S_{qq}(\omega)$ and $\hat{S}_{qq}(\omega)$ are positive definite for $\omega \in [-\pi, \pi]$. In the discussion, the predictive sample v will be considered to be conditional on $x(t)$

and to have an infinite sample size M . This will require the normalization of the negative entropy and I-divergence by the sample size M . The I-divergence per sample time conditional on $x(t)$, which will be denoted $I(S, \hat{S})$ and called *I-divergence* for brevity, can be expressed using (4) as (Kazakos and Papantoni-Kazakos, 1980)

$$\begin{aligned} I(S, \hat{S}) &= \lim_{M \rightarrow \infty} \frac{1}{M} K(p(v_M, \theta_0), p(v_M, \hat{\theta}(u_N))) \\ &= -\frac{1}{2} \int_{-\pi}^{\pi} \{\log |S_{qq}(\omega) \hat{S}_{qq}^{-1}(\omega)| + tr[I - S_{qq}(\omega) \hat{S}_{qq}^{-1}(\omega)]\} \frac{d\omega}{2\pi} \\ &\quad - \frac{1}{2} \int_{-\pi}^{\pi} tr\{\hat{S}_{qq}^{-1}[H(\omega) - \hat{H}(\omega)]S_{xx}(\omega)[H(\omega) - \hat{H}(\omega)]^T\} \frac{d\omega}{2\pi} \end{aligned} \quad (5)$$

where the subscript emphasizes that the sample of size M of v becomes infinite. The negative entropy per sample, or *negentropy* for brevity, is defined as $N(S, \hat{S}) = \lim_{M \rightarrow \infty} \frac{1}{M} R(p, \hat{p}) = E_v I(S, \hat{S})$. Note that the I-divergence is composed of two terms, the last due to the error in estimating the transfer function $H(\omega)$ and the first due to the error in estimating the spectrum $S_{qq}(\omega)$ of the noise $q(t)$. A useful approximation for the first term in (5) is

$$\begin{aligned} &-\frac{1}{2} \int_{-\pi}^{\pi} \{\log |S_{qq}(\omega) \hat{S}_{qq}^{-1}(\omega)| + tr[I - S_{qq}(\omega) \hat{S}_{qq}^{-1}(\omega)]\} \frac{d\omega}{2\pi} \\ &\approx \frac{1}{4} \int_{-\pi}^{\pi} tr\{S_{qq}^{-1}(\omega)[\hat{S}_{qq}(\omega) - S_{qq}(\omega)]^2\} \frac{d\omega}{2\pi} \end{aligned} \quad (6)$$

which holds to second order in the elements of \hat{S}_{qq} as is easily shown by comparing first and second derivatives of the integrands. This is a generalization to the multivariate case of the integral of the squared relative error. Thus the I-divergence is approximately a quadratic form in the estimation errors of $S_{qq}(\omega)$ and $H(\omega)$, and these quadratic forms do not interact, i.e. there are no cross terms.

3. NORMALIZED SPECTRAL ERROR IN PRINCIPAL COMPONENTS

In the multiple time series case, the spectral measure (5) has an intuitive interpretation in terms of principal components of the power spectrum in the frequency domain. Principal component representations of the spectral matrices $S_{qq}(\omega)$ and $S_{xx}(\omega)$ have the form

$$J(\omega) S_{qq}(\omega) J^*(\omega) = D(\omega) \quad , \quad L(\omega) S_{xx}(\omega) L^*(\omega) = E(\omega) \quad (7)$$

where $J(\omega)$ and $L(\omega)$ given as a function of frequency ω are unitary matrix transformations which diagonalize $S_{xx}(\omega)$ and $S_{yy}(\omega)$ respectively so that $J(\omega)J^*(\omega) = I = L(\omega)L^*(\omega)$, and where $D(\omega)$ and $E(\omega)$ are diagonal matrices.

Using spectral factorization theory, the matrix function $J(\omega)$ can be chosen as a continuous function of ω and the transfer function of a causal filter under either of the mild assumptions

- (i) The process is purely nondeterministic ((Gikhman and Skorokhod (1969), Whittle(1954) for the scalar random field case).
- (ii) The autocovariance function is absolute summable (Goodman and Ekstrom (1980) for the scalar random field case).

These derivations of the spectral factorization for the scalar case generalize to the multivariable case with care given to determining the logarithm of a matrix (Larimore, 1984, 1977). Orthonormalization of the rows of the spectral factor gives $J(\omega)$ while the normalizing terms form the diagonal of $D(\omega)$. Similarly, $L(\omega)$ can be taken as a spectral factor of a causal filter.

Let $X(\omega)$ be the random Fourier coefficients of $x(t)$, i.e. the spectral random measure of $x(t)$. Filtering $x(t)$ with transfer function $L(\omega)$ gives the principal component process $\bar{x}(t)$ which is expressed in the frequency domain as $\bar{X}(\omega) = L(\omega)X(\omega)$, and which has the diagonal spectral matrix $E(\omega)$ and similarly for $q(t)$.

Now consider the asymptotically equivalent expression (6) for the first term of the spectral measure (5) which is invariant to the unitary transformation $J(\omega)$ and is thus given by

$$\begin{aligned}
 & \frac{1}{4} \int_{-\pi}^{\pi} \text{tr} \{ D^{-1}(\omega) [\hat{D}(\omega) - D(\omega)]^2 \} \frac{d\omega}{2\pi} \\
 &= \frac{1}{4} \sum_i \int_{-\pi}^{\pi} \left[\frac{\hat{D}_{ii}(\omega) - D_{ii}(\omega)}{D_{ii}(\omega)} \right]^2 \frac{d\omega}{4\pi} + \frac{1}{2} \sum_{i \neq j} \int_{-\pi}^{\pi} \frac{|\hat{D}_{ij}(\omega)|^2}{D_{ii}(\omega) D_{jj}(\omega)} \frac{d\omega}{4\pi} \\
 &\approx \frac{1}{4} \sum_i \int_{-\pi}^{\pi} \left[\frac{\hat{D}_{ii}(\omega) - D_{ii}(\omega)}{\hat{D}_{ii}(\omega)} \right]^2 \frac{d\omega}{4\pi} + \frac{1}{2} \sum_{i \neq j} \int_{-\pi}^{\pi} \frac{|\hat{D}_{ij}(\omega)|^2}{\hat{D}_{ii}(\omega) \hat{D}_{jj}(\omega)} \frac{d\omega}{4\pi}
 \end{aligned} \tag{8}$$

where the approximation holds for the diagonal elements of $\hat{D}(\omega)$ near $D(\omega)$. The approximation is very useful when only the estimated spectrum $\hat{S}_{yy}(\omega)$ is known and we wish to consider the error in estimating the truth $S_{yy}(\omega)$. The first sum on the right hand side is the integrated squared relative error of the estimated cospectra of the principal components, while the second term is the integrated squared coherency of the estimated spectrum $\hat{D}(\omega)$ which would be zero if $\hat{D}(\omega) = D(\omega)$. Thus the first term of the measure (5) has a clear interpretation in the multivariate case when the true spectrum $D(\omega)$ is diagonal but where the approximating spectrum $\hat{D}(\omega)$ is

permitted

arbitrary coherency among components. The general case is reduced to this diagonal case by choosing an appropriate filter $J(\omega)$ which diagonalizes $S_{yy}(\omega)$ as in (7).

The second term in the spectral measure (5) is invariant to the unitary transformations $J(\omega)$ and $L(\omega)$ which gives

$$\begin{aligned}
 & -\frac{1}{2} \int_{-\pi}^{\pi} \{D^{-1}(\omega)[G(\omega) - \hat{G}(\omega)]E(\omega)[G(\omega) - \hat{G}(\omega)]^*\} \frac{d\omega}{2\pi} \\
 & = -\frac{1}{2} \int_{-\pi}^{\pi} \sum_{i,j} |\hat{G}_{ij}(\omega) - G_{ij}(\omega)|^2 \frac{D_{ii}}{E_{jj}} \frac{d\omega}{2\pi}
 \end{aligned} \tag{9}$$

where $G(\omega) = J(\omega)H(\omega)L^*(\omega)$ is the transfer function $H(\omega)$ expressed in the coordinate frame of the principal component series $\bar{x}(t)$ and $\bar{y}(t)$. The squared magnitude error $|\hat{G}_{ij}(\omega) - G_{ij}(\omega)|^2$ in the i,j element of the transfer function is weighted by the input signal to output noise ratio D_{ii}/E_{jj} for the pair i,j .

4. A LOWER BOUND ON ACHIEVABLE SPECTRAL ACCURACY

A lower bound on the expected negative entropy gives an asymptotic lower bound on the achievable accuracy in the estimation of the process spectrum and transfer function. This applies to the case of a fixed known model order as well as to the case of an unknown or infinite model order with the use of the AIC for model order selection. The achievable spectral accuracy is given as a function of the sample size and the number of parameters estimated.

Consider the case where the model \hat{S} is estimated using a K dimensional constrained estimator $\hat{\theta}^k$. As derived in Larimore (1986) using the Cramer-Rao lower bound, the expected negative entropy is asymptotically bounded by

$$E_{\mu} I(S, \hat{S}) \geq \frac{K}{2N} + E_{\mu} I(S, \bar{S}) \tag{10}$$

where \bar{S} is the model to which the constrained maximum likelihood estimate converges so that $E_{\mu} I(S, \bar{S})$ is the bias in the constrained model and $K/2N$ is the sampling variability.

The bound $K/2N$ is achieved for an asymptotically unbiased and efficient estimator such as maximum likelihood. In particular, this assumes that the true model order is no greater than the order K used in the model fitting. The true order K of the process is usually not known and may

in fact be infinite, so that the above discussion is not very useful in practice although it does give some insight into the accuracy issue.

Consider now the Akaike minimum AIC procedure (MAICE) using the estimator $\hat{\theta}_A$ (Akaike, 1973). Assume that the true model order is infinite, so that for any subset of the infinite parameter vector, there exist nonzero components. Thus it is not possible to obtain asymptotically unbiased estimates of θ using a fixed model order in estimating a model. Following Shibata (1983, 1981a, 1981b), define the optimal predictive order $K^*(N)$ depending upon the sample size N as the order K minimizing the negentropy (10). Then under suitable assumptions, the remarkable result is obtained by Shibata (1981) that asymptotically as $N \rightarrow \infty$

- (i) the lower bound using any order selection scheme is given by evaluating (10) at $K = K^*$, and
- (ii) this lower bound is achieved by the MAICE estimator $\hat{\theta}_A$.

Thus the lower bound on the negentropy which is achieved by MAICE is equal to the negentropy that would result from using an efficient estimator with apriori knowledge of the optimal order $K^*(N)$.

Using the spectral expression (5), an asymptotic lower bound on the expectation of the generalized relative squared error in estimating the power spectrum is given by

$$\begin{aligned} \frac{K^*(N)}{2N} \leq & \frac{1}{4} E_n \int_{-\pi}^{\pi} \text{tr} \{ S_{\eta\eta}^{-1}(\omega) [\hat{S}_{\eta\eta}(\omega) - S_{\eta\eta}(\omega)]^2 \} \frac{d\omega}{2\pi} \\ & + \frac{1}{2} E_n \int_{-\pi}^{\pi} \text{tr} \{ \hat{S}_{\eta\eta}^{-1} [H(\omega) - \hat{H}(\omega)] S_{xx}(\omega) [H(\omega) - \hat{H}(\omega)]^* \} \frac{d\omega}{2\pi} \end{aligned} \quad (11)$$

This gives a fundamental limit to the achievable accuracy in any parametric estimation procedure. A further perspective on this fact is given by the justification of the expected negentropy as the natural measure of modeling approximation error in statistical inference.

REFERENCES

- Akaike, H., (1973). "Information theory and an extension of the maximum likelihood principle." In *2nd International Symposium on Information Theory*, Eds. B.N. Petrov and F. Csaki, pp. 267-281. Budapest: Akademiai Kiado.

Gikhman, I.I. and A.V. Skorokhod (1969). *Introduction to the Theory of Random Processes*. Philadelphia: Saunders Company.

Goodman, D.M. and M.P. Ekstrom (1980). "Multidimensional spectral factorization and unilateral autoregressive models." *IEEE Trans. Automatic Control*, Vol. 25, pp. 258-62.

Kazakos, D., and P. Papantino-Kazakos (1980). "Spectral Distance Measures Between Gaussian Processes," *IEEE Trans. Automat. Control*, Vol. 25, pp.850-59.

Larimore, W.E. (1986). "Constrained Nonnested Multiple Comparison Using Predictive Inference and Entropy," Draft.

Larimore, W.E. (1984). Efficient Computation of Maximum Likelihood Estimates for Stochastic Space-Time Models with Incomplete Data, Final Technical Report, prepared under Contract No. NASO-82-ABC-00241 for NOAA/National Marine Fisheries Service. Scientific Systems, Inc, Cambridge, MA.

Larimore, W.E. (1977). "Statistical inference on stationary random fields." *Proc. IEEE*, Vol 65 , pp. 961-70.

Shibata, R., (1983). "A Theoretical View of the Use of AIC," *Proc. of the International Time Series Meeting*, O.D. Anderson, Ed.

Shibata, R., (1981a). "An optimal autoregressive spectral estimate." *Ann. Statist.* Vol 9, pp. 300-6.

Shibata, R., (1981b). "An optimal selection of regression variables." *Biometrika*, Vol 68, pp. 45-54.

Whittle, P. (1954). "On stationary processes in the plane," *Biometrika*, Vol 41, pp. 433-49.

END

11-87

DTIC